



UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

Smoothing mixed models for spatial and spatio-temporal data

Autor:
Dae-Jin Lee

Directora:
María L. Durbán Reguera

DEPARTAMENTO DE ESTADÍSTICA

Leganés, mayo 2010

TESIS DOCTORAL



Universidad
Carlos III de Madrid

Ph.D. Thesis

**Smoothing mixed models for spatial and
spatio-temporal data**

Author:

Dae-Jin Lee

Advisor:

María L. Durbán Reguera

DEPARTMENT OF STATISTICS

Leganés, May 2010

© 2010
Dae-Jin Lee
All Rights Reserved

To my parents

Acknowledgements

I want to express my gratitude to my supervisor Dr. María Durbán, who has provided her support in numerous ways, who shared with me a lot of her expertise and research insight, and whose encouragement, guidance, friendship and support from the beginning to the final stage enabled me to write this thesis.

I am very grateful to the Department of Statistics of Universidad Carlos III de Madrid, for giving me the financial support and where I could develop my work as teaching assistant.

I would like to give a special mention to all the colleagues, class and office mates I had lucky chance to meet during the last years at the Department of Statistics at the Universidad Carlos III de Madrid in Leganés, thanks to: Peter, Bernardo, Emilio, Nacho, Sergio, Elisa, Ismael, Javier N., Ángel, Sofi, Isaac, Vicente, and Mari. I would also like to extend this mention to the people I had met at the IWSM conferences I had attended since 2006 in Galway: Iain, Paul, Jutta, and Giancarlo.

I wish to thank Prof. Adrian W. Bowman for his kindness and hospitality during the time I spent at the Department of Statistics at the University of Glasgow.

To my parents, who came from South Korea to Spain, and gave my older brother and me all the support to get an education and who taught us to be, first of all, honest people. Probably, only I will understand all the magnitude of their sacrifice when I have my own family.

Finally, to my friends, my other *familee*.

I made it after all!

Gracias a todos

The research presented in this thesis was supported by the research projects: Comunidad de Madrid (CCG06-UC3M/ESP-0856), and Spanish Ministry of Science and Innovation (projects SEJ2005-06454 and MTM2008-02901).

Abstract

The development of many of statistical methods and models has been linked to the study of specific applications within various scientific research fields. The analysis of spatial and spatio-temporal data is currently of great interest to statistical modelling. Problems related to meteorology, environmental pollution, ecology, epidemiology or economics, demand the use of statistical models for spatial and spatio-temporal data. In the first chapter of this thesis, we introduce the basic concepts in spatial statistics, the classification of spatial data according to their typology and a review of the classical models in the literature and their limitations.

In this thesis, we propose the modelling of these data using non-parametric regression methods, also known as *smoothing techniques*. Our proposal is to consider the modelling from an unified perspective for the different types of spatial data, by means of the use of so-called penalized splines (*P-splines*). These models have become very popular in recent years as: (i) they are *low-rank smoothers*, because they are constructed from regression basis (*B-splines*) of smaller dimension than the number of observations, so they are computationally more efficient than other splines-based methods; (ii) the formulation as a mixed model allows the incorporation of more complex structures in terms of random effects that are estimated simultaneously to the smoothing. The second chapter is entirely dedicated to introducing the fundamentals of the *P-spline* methodology for Gaussian data and in the context of generalized linear models for non-Gaussian data. In multidimensional problems, the regression basis is defined as the Tensor product of the marginal *B-spline* bases, that in the case of data in regular multidimensional grids it is the Kronecker product of matrices of *B-splines*. For these situations, the array methods allow to fit the models in a computationally efficient way. We also detail the representation as a mixed model and estimation methods. Although this representation is not new in the literature of splines, our reparameterization of the basis and penalty of the model allows the decomposition of the fit in terms of the sum of marginal functions and interactions. Finally, in this chapter, we adapt the array algorithms to mixed models.

In the third chapter, we extend the *P-spline* models for smoothing spatial data. The

structure of spatial data requires the use of a new Tensor product, the *row-wise* Kronecker product. Then for this particular case, the array methods are not applicable, and the mixed model reparameterization is not immediate, however, we demonstrate how it can be obtained using some matrix algebra results. We illustrate the methodology for the types of spatial data and examples shown in chapter one. As an application of the methodology proposed in this chapter, we discuss the analysis of regional count data. Count data are usually assumed distributed according to a Poisson random variable, however, this assumption is sometimes incorrect when the data have an unexplained heterogeneous variability (*overdispersion*). As a results of this chapter, in Lee and Durbán (2009), we analyzed the well-known scottish lip cancer data. These data have been widely used in the literature of models for regional data, especially from the conditionally autoregressive (CAR) models approach. We propose a *hybrid* smooth model that allows to incorporate different sources of spatial variability: (i) a *large-scale* spatial variability, captured by the spline, and (ii) a local *small-scale* variability defined by the neighborhood structure of the regions of the study, with a CAR structure. The advantage of our hybrid model is that both sources can estimate simultaneously. The simulation studies carried out confirm that the hybrid model can capture the different sources of variability in the proposed scenarios.

In the fourth chapter, we consider the multidimensional case by decomposing the model as the sum of smooth functions in terms of main or additive effects and interactions (these models are called Smooth-ANOVA models, by analogy to the factorial design and analysis-of-variance). The construction of these models with *B*-spline bases, suffers from problems of identifiability since they cannot be estimated uniquely. Our solution to avoid these problems is the reparameterization as mixed models developed in previous chapters. This reparameterization allows us to identify what are the elements that are repeated, the solution to the problem is then reduced to eliminate the repeated terms. This procedure allows a simple way to build a new basis and penalty for the identifiable model. The interesting result of this simple procedure is exactly equivalent to apply linear constraints to the regression coefficients of the original model. The simulation study presented in this chapter, shows that the Smooth-ANOVA model performs in the same way as the most appropriate model for each of the scenarios considered. In some situations, it is of interest to consider in the modelling only some main effects and interactions and ignore the rest. These models, models are called *reduced* Smooth-ANOVA models. An example of this is the spatio-temporal case, where this decomposition allows to represent the smoothing in terms of the sum of a spatial surface, a smooth function for the temporal component and a smooth term for the space-time interaction. For the spatio-temporal case, we construct the regression basis as the Kro-

necker product of the B -spline bases of space and time dimensions. This allows us to use the array methods defined in previous chapters. Following the procedure described in this chapter, we show how to construct the models and identify the linear restrictions on the regression coefficients of the original model. In Lee and Durbán (2010), we apply the reduced S-ANOVA model to the spatio-temporal analysis of ozone levels in Europe during period 1999-2005. Finally, in this chapter we propose a computationally efficient methods for models with interactions. In some cases, the size of the B -spline basis for the interaction is very large, which implies that the parameter estimation is computationally intensive. For the case of Smooth-ANOVA models, it is possible to assume that most of the structure is captured by the main effects, and is therefore preferable to reduce the complexity of the model by reducing the size of the bases of the interaction. However, this reduction is not arbitrary, since otherwise the models would not be nested. Our proposal is the construction of nested B -spline basis for the interactions. For the spatio-temporal case, this solution allows to model the temporal part with a larger basis to capture the time structure of the data, and a nested basis (much smaller) for the space-time interaction.

Finally, the fifth chapter summarizes the main contributions made in this thesis, we suggest possible future extensions to the models developed and new lines of research.

Resumen

El desarrollo de gran parte de los modelos y métodos estadísticos ha ido ligado al deseo de estudiar aplicaciones específicas dentro de diversos ámbitos científicos. El análisis de datos de naturaleza espacial y espacio-temporal es en la actualidad de gran interés para la modelización estadística. Problemas relacionados con la meteorología, la contaminación medioambiental, la ecología, la epidemiología o la economía, demandan el uso de modelos estadísticos para el análisis de datos espaciales y espacio-temporales. En el primer capítulo de esta tesis, introducimos los conceptos básicos de la estadística espacial, así como la clasificación de los datos espaciales según su tipología y una revisión de los modelos tradicionales en la literatura con sus principales limitaciones.

En esta tesis proponemos la modelización de este tipo de datos mediante modelos de regresión no-paramétricos, también denominadas *técnicas de suavizado*. Nuestra propuesta es considerar la modelización desde una perspectiva común para los diferentes tipos de datos espaciales, mediante el uso de los denominados *splines* con penalizaciones (*P-splines*). Estos modelos han adquirido una gran popularidad en los últimos años ya que: (i) se tratan de *suavizadores* de rango bajo, ya que se construyen a partir de *bases* para la regresión de (*B-splines*) menor tamaño que el número de datos, que son computacionalmente más eficientes que otros métodos de suavizado basados en *splines*; (ii) la formulación como modelos mixtos permite incorporar estructuras más complejas en términos de efectos aleatorios que pueden estimarse simultáneamente al suavizado. El segundo capítulo está enteramente dedicado a introducir los aspectos fundamentales de la metodología de los *P-splines*, para datos Gaussianos y en el contexto de los modelos lineales generalizados para el caso de datos no-Gaussianos. Para el caso multidimensional, la base para la regresión se define como el producto Tensorial de las bases de *B-spline* marginales, que en el caso de datos en *grids* o mallas multidimensionales es el producto de Kronecker de las matrices de *B-spline*. En esta situación, el uso de los métodos de *array* permite el ajuste de los modelos de manera computacionalmente eficiente. Presentamos también en detalle la representación como modelo mixto, y los métodos de estimación. Aunque esta representación no es nueva en la literatura de los *splines*, nuestra reparametrización de las bases y de la penalización del modelo

permite la decomposición del ajuste en términos de la suma de funciones marginales e interacciones. Por último en este capítulo, adaptamos los algoritmos basados en arrays para la formulación como modelo mixto.

En el tercer capítulo, extendemos los modelos de P -splines para el suavizado de datos espaciales. La estructura de los datos espaciales requiere del uso de un nuevo producto Tensorial, el producto de Kronecker *por filas*. En este caso, los métodos de arrays no son aplicables, y la reparametrización como modelo mixto no es inmediata, sin embargo demostramos cómo se puede llegar a ella mediante resultados matriciales. Ilustraremos la metodología para la tipología de datos y ejemplos de datos espaciales introducidos en el primer capítulo. Como aplicación de la metodología propuesta en este capítulo, abordamos el análisis de datos regionales de conteo. Los datos de conteo se asumen distribuidos según una variable aleatoria Poisson, sin embargo, este supuesto resulta en ocasiones erróneo cuando los datos presentan una variabilidad heterogénea no explicada (*sobre-dispersión*). Como resultado de este capítulo, en Lee y Durbán (2009), analizamos los datos de cáncer de labio en Escocia. Estos datos han sido muy utilizados en la literatura de los modelos para datos regionales, sobre todo desde el enfoque de los modelos condicionalmente autorregresivos (CAR). En este trabajo proponemos modelos de suavizado *híbridos* que permiten incorporar diferentes fuentes de variabilidad espacial: (i) una variabilidad espacial a *gran escala*, capturada por el spline, y (ii) una variabilidad local a *pequeña escala* dada por la estructura de vecindad de las regiones del estudio, con una estructura tipo CAR. La ventaja de nuestro modelo híbrido, es que ambas fuentes se pueden estimar simultáneamente. Los estudios de simulación realizados corroboran que el modelo híbrido permite capturar las diferentes fuentes de variabilidad en los escenarios propuestos.

En el cuarto capítulo, consideramos el caso multidimensional mediante la descomposición de los modelos como la suma de funciones de suavizado, en términos de efectos principales o aditivos e interacciones (estos modelos son denominados modelos de suavizado ANOVA, por su analogía a los diseños factoriales y análisis de la varianza). La construcción de estos modelos mediante bases de B -spline, sufre de problemas de identificabilidad dado que no se pueden estimar de manera única. Nuestra solución para evitar estos problemas es la reparametrización como modelos mixto desarrollada en los capítulos anteriores. Esta reparametrización permite identificar cuáles son los elementos que aparecen repetidos, la solución al problema se reduce por tanto a eliminar los componentes repetidos, lo cual permite de manera sencilla construir la nueva base y la penalización para el modelo identificable. Lo interesante de este sencillo procedimiento es su equivalencia a imponer restricciones lineales sobre los coeficientes del modelo original. El estudio de simulación presentado en este capítulo, demuestra que el

modelo de suavizado ANOVA actúa bajo los escenarios considerados del mismo modo que el modelo más apropiado para cada caso. En algunas situaciones, resulta de interés considerar tan sólo algunos efectos principales e interacciones e ignorar otros. Estos modelos, reciben el nombre de modelos *reducidos* de suavizado ANOVA. Un ejemplo es el caso espacio-temporal, donde resulta de interés la decomposición del proceso en términos de la suma de una superficie espacial, una función suave para el componente temporal, y un componente espacio-temporal que recoge la interacción espacio-tiempo. Para el caso espacio-temporal, construiremos las bases para la regresión mediante el producto de Kronecker de las bases de B -spline espacial y temporal, lo cual permite para este caso utilizar los métodos de array definidos en los capítulos anteriores. Siguiendo el procedimiento desarrollado en este capítulo, demostramos cómo construir los modelos e identificamos las restricciones sobre los coeficientes en el modelo original. Para ilustrar esta metodología en Lee y Durbán (2010), consideramos el uso de estos modelos para el análisis espacio-temporal de los niveles de ozono en Europa entre los años 1999 y 2005. Por último, en este capítulo, proponemos un método computacionalmente eficiente para el caso de modelos con interacción. En algunas situaciones, el tamaño de la matriz de B -splines para la interacción es muy grande, lo que conlleva a que la estimación de los parámetros sea computacionalmente intensiva. En el caso de los modelos de suavizado ANOVA, es posible asumir que la mayor parte de la estructura es recogida por los efectos principales, y por tanto es preferible reducir la complejidad del modelo reduciendo el tamaño de las bases de la interacción. Sin embargo, esta reducción no es arbitraria, puesto que de otro modo los modelos no estarían anidados. Nuestra propuesta es la construcción de *bases anidadas* de B -spline para las interacciones. En el caso espacio-temporal, esta solución permite modelizar la parte temporal con una base de más tamaño para recoger la estructura temporal de los datos, y una base anidada (mucho más pequeña), para modelizar la interacción espacio-tiempo.

Finalmente, en el quinto capítulo resumimos las principales aportaciones realizadas en esta tesis, y proponemos posibles futuras extensiones a los modelos desarrollados y nuevas líneas de investigación.

Contents

List of Figures	xiii
1 Spatial statistics: data and models	1
1.1 Spatial data analysis: definitions and examples	2
1.1.1 Geostatistical data	2
1.1.2 Regional or areal data	3
1.1.3 Point patterns	4
1.2 Classic models for the analysis of spatial data	5
1.2.1 Kriging methods	5
1.2.2 Spatial models for regional data	9
1.2.3 Point patterns models	10
1.2.4 Hierarchical spatial models	14
1.3 Spatio-temporal modelling	15
1.4 The smoothing approach	18
2 Smoothing mixed models	23
2.1 Penalized splines: an introduction	23
2.1.1 Bases and penalties	24
2.1.2 Some basic definitions	28
2.1.3 Smoothing parameter selection	30
2.1.4 P -splines for non-Gaussian responses	31
2.2 Penalized splines and mixed models	35
2.2.1 Mixed models representation of P -splines	37
2.2.2 P -splines as generalized linear mixed models	42
2.3 Multidimensional smoothing with P -splines	43
2.3.1 Smoothing multidimensional data with array structure	44
2.3.2 Multidimensional mixed models representation of P -splines	48
2.3.3 Array methods for multidimensional smoothing mixed models	56

3	Smoothing spatial data with penalized splines	61
3.1	<i>B</i> -spline basis for spatial data	61
3.2	Mixed model reparameterization for spatial data	63
3.3	Examples of smoothing spatial data with <i>P</i> -splines	65
3.4	Smoothing mixed models for spatial count data	70
3.4.1	Overdispersion in Poisson count data	70
3.4.2	Spatial smoothing mixed models with CAR structure	76
3.5	Simulation study	84
4	Smooth-ANOVA models	95
4.1	<i>P</i> -spline additive models	96
4.1.1	Smoothing additive mixed models	97
4.2	<i>P</i> -spline smooth-ANOVA models	101
4.2.1	Smoothing additive mixed models with interactions	101
4.2.2	Reparametization of the S-ANOVA model into a mixed model for- mulation	103
4.2.3	Transformation matrix in S-ANOVA models	104
4.2.4	Smooth-ANOVA models construction	109
4.2.5	Simulation of smooth surfaces	111
4.3	Testing components in smoothing mixed models	115
4.4	Reduced S-ANOVA models for spatio-temporal data	118
4.4.1	Spatio-temporal <i>P</i> -spline models and basis	118
4.4.2	Transformation matrix in the reduced spatio-temporal S-ANOVA model	121
4.4.3	Linear constraints over coefficients in the reduced spatio-temporal S-ANOVA model	125
4.4.4	Analysis of air pollution levels in Europe	130
4.5	Smooth-ANOVA models and nested <i>B</i> -spline bases	133
4.5.1	Nested <i>B</i> -spline basis for spatio-temporal data	136
4.6	Further considerations	140
5	Conclusions and further work	145
	References	151
A	Appendix to Chapter 2	165
A.1	Some basic matrix algebra on Kronecker products	165
A.2	Array methods	168
A.2.1	Basic array arithmetic	168

A.2.2	GLAM algebraic operations	170
A.2.3	GLAM as mixed models	171
A.3	Software considerations	173
A.3.1	Function <code>lme()</code> in <code>nlme</code> R package	174
A.3.2	Function <code>glmmPQL()</code> in <code>MASS</code> R package	176
B	Appendix to Chapter 3	177
B.1	On matrix algebra of Khatri-Rao and Tracy-Singh products	177

List of Figures

1.1	Sample of monitoring stations across Europe. Source: European monitoring and evaluation programme. webpage: http://www.eea.europa.eu/	2
1.2	Sudden-infant-death syndrom (SIDS) counts in North Carolina in 1974. . .	3
1.3	SIDS data contiguity with four nearest neighbors criteria	4
1.4	Locations of trees in Lansing Woods divided by their botanical classification. The spatial locations have been rescaled to the unit square.	5
1.5	Matèrn correlation functions, for different values of ν and range parameter ϕ	8
1.6	Kernel estimation in spatial point patterns	12
1.7	Estimated kernel density estimates of maple trees from the Lansing data set with Gaussian kernel and $h = \{0.1, 0.08, 0.06, 0.02\}$	13
1.8	(a) sample of 43 monitoring stations over Europe. (b) O_3 levels in four selected countries.	16
2.1	B -spline regression bases of different orders of degree p and $m = 6$ and equally-spaced knots	26
2.2	(a) fitted curve with unpenalized coefficients (red circles). Bottom: fitted curve with penalized coefficients (blue circles).	28
2.3	Fitted P -spline curves with different values of $\lambda = \{10^{-6}, 1, 10^3, 10^6\}$ and a second order penalty, $d = 2$	29
2.4	Fitted Poisson P -GLM to Greeks data, with optimized $\lambda = 39.81$ by BIC. .	35
2.5	Tensor product of two cubic splines.	45
2.6	A portion of the full basis, consisting in the Tensor product of nine cubic splines.	45
2.7	Raw data and smoothed data with 8-by-5 knots and	47
2.8	Decomposition of the two dimensional surface into additive terms and interactions.	53
2.9	Array Θ of coefficients in three dimensions of $c_1 \times c_2 \times c_3$	55

3.1	Smoothed surface of O_3 levels in January 1999.	66
3.2	Fitted smooth trend of SIDS data in 1974, using the centroids of the counties as spatial locations.	67
3.3	2d histograms of counts of maples trees with different number of bins in each direction, $n_{bins} = \{15, 30, 60, 100\}$	68
3.4	Smoothed intensity functions for different number of bins.	69
3.5	Comparison of fitted curves for smoothing mixed models: Poisson, PRIDE and Negative Binomial	75
3.6	Neighboring structure for Scottish data: (a) Contiguity defined in Breslow and Clayton (1993); (b) contiguity based on sharing a common border.	80
3.7	CAR model: (a) Linear Trend ($X\beta$); (b) CAR random effect (b) with G_b defined by Dean in (3.33) and (c) CAR model fit ($X\beta + b$).	81
3.8	PRIDE Model: (a) Spatial Smooth Trend ($X\beta + Z\alpha$); (b) Overdispersion individual random effects (γ) and (c) the sum of trend and overdispersion effects.	82
3.9	Smooth-CAR model: (a) Smooth Trend ($X\beta + Z\alpha$); (b) CAR structured random effects (b) and (c) the sum of trend and CAR component.	83
3.10	(a) Spatial deviance residuals for fitted models and (b) locations of regions of Scotland with larger values of the residuals.	83
3.11	log(MSE) comparison of Poisson, PRIDE, Smooth-CAR and CAR models in scenario 2 with $R = 100$	86
3.12	log(MSE) comparison of Poisson, PRIDE, Smooth-CAR and CAR models in scenario 2 with $R = 100$	87
3.13	log(MSE) comparison of PRIDE, Smooth-CAR and CAR models in scenario 3 with $R = 100$ and $\sigma_s = 0.25$	89
3.14	log(MSE) comparison of PRIDE, Smooth-CAR and CAR models in scenario 3 with $R = 100$ and $\sigma_s = 1$	89
3.15	log(MSE) comparison of PRIDE, Smooth-CAR and CAR models in scenario 4 with $R = 100$ and $\sigma_s = 0.25$	92
3.16	log(MSE) comparison of PRIDE, Smooth-CAR and CAR models in scenario 4 with $R = 100$ and $\sigma_s = 0.75$	92
4.1	Simulated functions: (a) and (b) are the nonlinear main effects of x_1 and x_2 ; (c) is the additive surface of main effects; (d) is interaction surface and (e) is the sum of the main effects and the interaction surfaces.	112
4.2	log(MSE) of fitted smooth models in scenario 1 and $R = 200$	113
4.3	log(MSE) of fitted smooth models in scenario 2 and $R = 200$	113
4.4	log(MSE) of fitted smooth models in scenario 3 and $R = 200$	113

4.5	Array $\Theta^{(\text{st})}$ of coefficients for the space-time interaction, of dimensions $c_t \times c_1 \times c_2$.	127
4.6	Restrictions over the array $\Theta^{(\text{st})}$, in spatial dimensions	129
4.7	Spatial and temporal smooth terms for S-ANOVA model.	132
4.8	Spatio-temporal interaction fit for the spatio-temporal S-ANOVA model, from March to August 2002.	134
4.9	Comparison of fitted values for monitoring stations in Spain, Sweden, Austria and UK.	135
4.10	Illustrative example of two nested B -spline bases, with $d = 2$.	137
4.11	U.S. monthly average temperature (in $^{\circ}\text{F}$) data of 136 cities from January 1995 and December 2004 ($t = 120$ time points). The total number of observations is 16320.	138
4.12	U.S. temperature data time trend: $f_t(\mathbf{x}_t)$. Fitted with 30,20,15 and 10 knots in the construction of the B -spline basis \mathbf{B}_t , in the reduced S-ANOVA without nested basis in the space-time interaction.	139
4.13	U.S. temperature data spatial effect: $f_s(\mathbf{x}_1, \mathbf{x}_2)$. The Figure shows the a south to north spatial pattern.	139
4.14	Sine function: $\beta_0 + \gamma \sin(2\pi(\mathbf{x}_t - \varphi)/p)$.	141
4.15	Comparison of standard and cyclic cubic B -spline basis. Both figures represent the first four columns of the cyclic B -spline basis.	142

*"Don't try to be original,
just try to be good".
Paul Rand*

Chapter 1

Spatial statistics: data and models

In last decades, *spatial statistics* has become an emerging area of research in many different fields. Spatial data arises from diverse fields as ecology, environmental sciences, epidemiology, geography, sociology or economics. The effort of studying spatial data from such diverse areas, has led to a wide variety of different approaches. From a statistical point of view, spatial data are realizations of random variables collected in geographical locations. Modern spatial statistical approaches are challenged to absorb and combine the wide variety of tools and concepts and develop new mathematical models in order to provide an useful explanation to the underlying spatial phenomema.

The main aim of the statistical spatial models are not only to quantify the information of the collected data, but to answer questions like: *how* and *where*. The recent proliferation of geographical information systems (GIS) software contributes to the development of new techniques. Because of spatial data arise in diverse fields and applications, there exists a variety of spatial data types, modelling approaches, and scenarios. Spatial data are also collected over several time periods, this incorporates a new dimension on the modelling, since the spatial process (in most of the cases) changes over time. Unfortunately, statistical tools for the analysis of *spatio-temporal* processes are not fully developed and more sophisticated tools are usually required.

This Chapter is concerned to the basic definitions, types of data and most traditional modelling approaches considered in the spatial statistics literature. In [Section 1.1](#), we present the classification of spatial data used in the literature. [Section 1.2](#), addresses the variety of methodological aspects of the different approaches considered for the spatial data classification. In [Section 1.3](#), we present the extension to the spatio-temporal case. Finally, [Section 1.4](#) introduces the *smoothing* approach we take throughout the thesis.

1.1 Spatial data analysis: definitions and examples

The main characteristic in the analysis of spatial data is the presence of a spatial dependence or autocorrelation among observations in the space. Observations in close locations are expected to be more similar than those that are more spatially separated. The purpose of a spatial model is to be able to describe the spatial variation across the surface of study.

A *spatial process* is defined as the random variable $Y(s)$:

$$\{Y(s) : s \in \mathcal{D}\},$$

where s indicates the location of the spatial observation in a d -dimensional space, where s varies over a *domain*, i.e. $\mathcal{D} \subset \mathbb{R}^d$. In general, we consider a two-dimensional space, $d = 2$, where $s = (x_1, x_2)$ are the geographical coordinates (longitude and latitude).

The classification of spatial data is the first step to specify which modelling approach is preferable. We adopt the classification proposed by Cressie (1993), that divides the class of spatial data according to the nature of the domain \mathcal{D} . Cressie (1993) classifies spatial data as: *geostatistical*, *regional* or *lattice data* and *point patterns*.

1.1.1 Geostatistical data

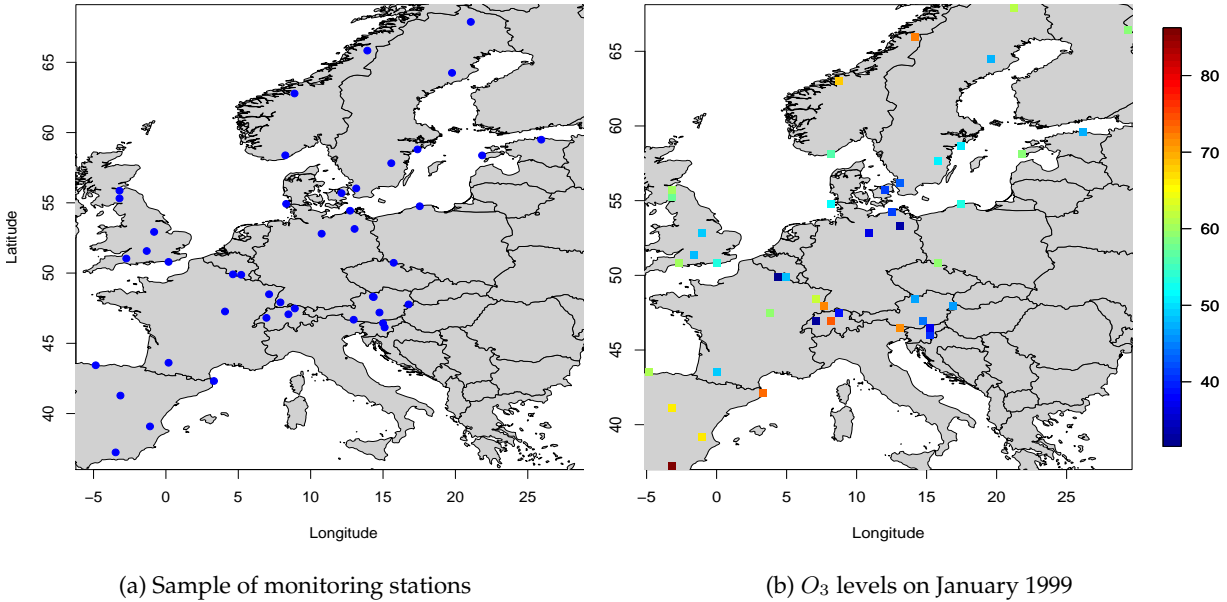


Figure 1.1: Sample of monitoring stations across Europe. Source: European monitoring and evaluation programme. webpage: <http://www.eea.europa.eu/>

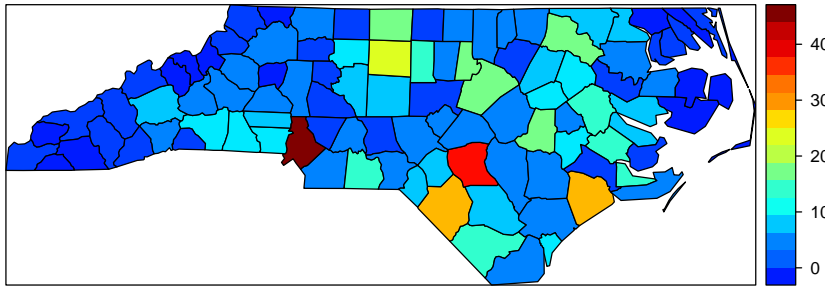


Figure 1.2: Sudden-infant-death syndrome (SIDS) counts in North Carolina in 1974.

Geostatistical data are characterized by a fixed and continuous domain \mathcal{D} . The spatial process $Y(s)$, can be observed in any continuous locations in \mathcal{D} , and then $Y(s)$ is a random variable in each of the spatial locations $s \in \mathcal{D}$. In the analysis of geostatistical data, the aim is to predict values of the attribute considered in locations where data are not available, or to reconstruct a surface of the attribute Y over the entire domain \mathcal{D} .

Figure 1.1 shows an example of geostatistical data. The European Environmental Agency (EEA), is an agency of the European Union involved in implementation and development of environmental policies. Figure 1.1a presents a sample of monitoring stations located across Europe. These monitoring stations are the fixed locations at geographical longitude and latitude, that take measurements of several environmental attributes as temperature or levels of pollutants. Figure 1.1b consider the measurements on Ozone (O_3). A geostatistical model, would be able to provide a mathematical model to study the spatial trend of O_3 levels and to predict the levels of O_3 at locations where no measurements were collected.

1.1.2 Regional or areal data

Regional or areal data are spatial data where the domain \mathcal{D} is a fixed and discrete set of points. Each of these spatial points are indexed such that $s_i \in \mathcal{D}$, $i = 1, 2, \dots, n$. Each spatial point or *lattice* can be of regular or irregular shape, and are often referred to as district levels, areas, regions or countries. Figure 1.2 is an example of regional data. This data set was studied by Symons et al. (1983), and from a spatial analysis in Cressie and Read (1985) and in more details in Cressie (1993). The data consists in the analysis of sudden-infant-death (SID) counts in the 100 counties of North Carolina, USA. In regional data, the spatial location s_i corresponds to a geographic region.

In most situations, regional data are spatially aggregated as events counts (e.g. num-

ber of deaths). Then, it results very common to assume that areas in close proximity to another are spatially correlated. The spatial structure is commonly based on the idea of *spatial connectivity*: those locations that are spatially connected are considered as *neighbors*. Another possibility is to consider the Euclidean distance between the centroids of each region, and then, the regional data can be viewed as geostatistical data.

Spatial connectivity can be defined by a matrix $W = \{\omega_{ij}\}$, with

$$W = \{\omega_{ij}\} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ are connected, } i \neq j \\ 0 & \text{otherwise.} \end{cases} \quad (1.1)$$

There exist several neighborhood criteria, for instance, the distance between regions, k -nearest neighbors or sharing a common boundary. An example of this criteria is shown in Figure 1.3, where we show the spatial connectivity of the 100 counties of North Carolina, US, using the four nearest neighbors criteria.

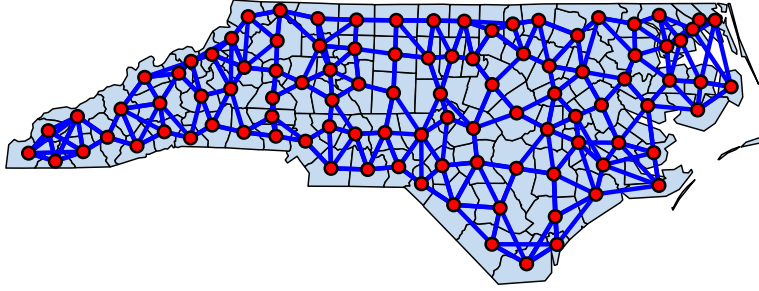


Figure 1.3: SIDS data contiguity with four nearest neighbors criteria

1.1.3 Point patterns

In geostatistical and areal data, the domain \mathcal{D} is fixed, this means that \mathcal{D} does not change from one realization to another. In spatial point patterns, the spatial domain \mathcal{D} is itself random, and it is a collection of points where an event has occurred. This type of spatial data are usually defined as:

$$Y(\mathbf{s}) = \begin{cases} 1 & \text{for all } \mathbf{s} \in \mathcal{D} \\ 0 & \text{otherwise.} \end{cases}$$

Spatial point pattern data are similar to geostatistical data if $Y(\mathbf{s})$ is binary. The main interest of the study of point patterns are the spatial locations itself, to determine if the

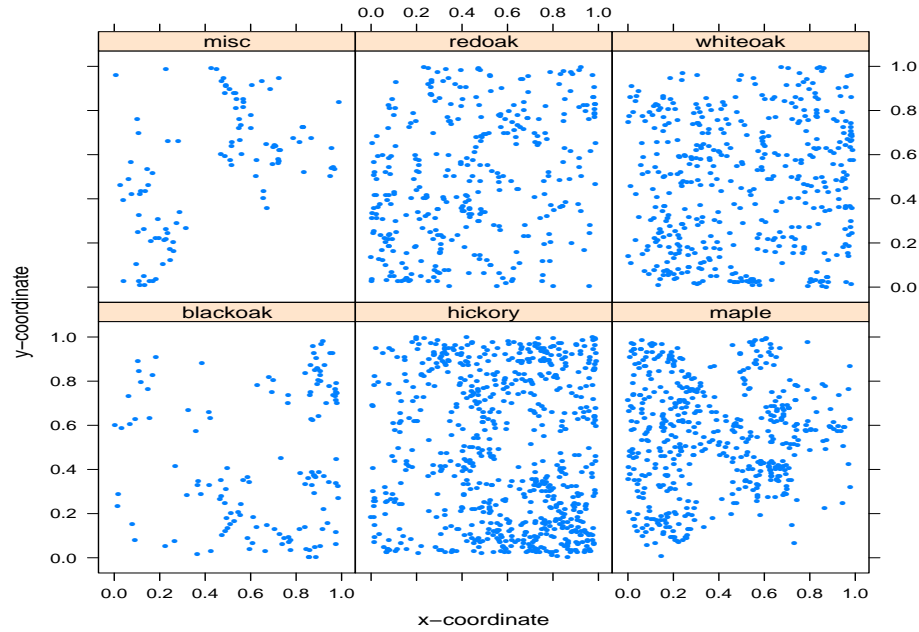


Figure 1.4: Locations of trees in Lansing Woods divided by their botanical classification. The spatial locations have been rescaled to the unit square.

events present a clustering pattern (the events occur close to others), or if the events are spatially random or independent. Figure 1.4 presents data from an investigation in Lansing Woods, Clinton County, Michigan USA. The data set consists of the locations of 2251 trees and their botanical classification (into hickories, maples, red oaks, white oaks, black oaks and miscellaneous trees). A question of interest is to study the spatial pattern of the concentration of the trees species.

1.2 Classic models for the analysis of spatial data

In this Section we present the most commonly used models in the spatial statistics literature according to the spatial data classification seen in Section 1.1. The main aim of this Section is to present a summary of classic methods to point out the heterogeneity of the existing modelling approaches and the assumptions adopted for each class of spatial data types.

1.2.1 Kriging methods

Kriging are a family of geostatistical models for the interpolation of geostatistical spatial data by generalized least squares regression techniques (Krige, 1951; Matheron, 1962,

1963). Let $y(s_i)$ be the set of observed values of the random variable Y , at locations s_i , for $i = 1, 2, \dots, n$. A general kriging model for the data is defined as:

$$y(s_i) = Z(s_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad (1.2)$$

where $\{Z(s) : s \in \mathbb{R}^d\}$. The interpolation at a new location s_0 , is done by:

$$\hat{Z}(s_0) = \sum_{i=1}^n \omega_i(s) Z(s_i) = \omega' Z(s_i), \quad (1.3)$$

where $\omega = (\omega_1(s), \dots, \omega_n(s))$ is the vector of kriging weights that are calculated such that $\hat{Z}(s_0)$ is the best linear predictor (BLP) of $Z(s_0)$. Then, $\hat{Z}(s_0)$ is unbiased and has minimum variance. The kriging weights are then:

$$\hat{\omega} = C^{-1}c_0,$$

where

$$C = \text{Cov} \{ [Z(s_1), \dots, Z(s_n)]' \} \text{ and } c_0 = [\text{Cov}\{Z(s_0), Z(s_1)\}, \dots, \text{Cov}\{Z(s_0), Z(s_n)\}]'.$$

The estimation is done by generalized least squares, which gives the solution:

$$\hat{Z}(s_0) = c_0'(C + \sigma^2 I)^{-1} y. \quad (1.4)$$

The equation in (1.4) indicates that all we need is the covariance structure of Z , to obtain the BLP. Kriging methods are based on two assumptions: stationarity, and isotropy.

Stationarity

Let be $\mu(s)$, the mean of the spatial process, such that $\mathbb{E}[Y(s)] = \mu(s)$. The spatial process is *strictly stationary* if for any given set of locations $\{s_1, \dots, s_n\}$, and $h \in \mathbb{R}^d$, the distribution of $(Z(s_1), \dots, Z(s_n))$ is the same as the distribution of $(Z(s_1 + h), \dots, Z(s_n + h))$. Cressie (1993) defines a less restrictive concept of stationarity (*weak* or *second order stationarity*), implying that the spatial process $Z(s)$ has a constant mean $\mu(s) = \mu$, and covariance function defined by:

$$\text{Cov}[Z(s), Z(s + h)] = \mathcal{C}(h), \text{ for all } h \in \mathbb{R}^d.$$

Note that, weak stationarity implies that the covariance between the values of the spatial process at any given two different locations are expressed by the covariance function

Table 1.1: Most common type of Matérn covariance functions.

ν	$\rho(r, \phi, \nu)$
1/2	$\exp(- r/\phi)$
3/2	$\exp(- r/\phi)(1 + r/\phi)$
5/2	$\exp(- r/\phi)(1 + r/\phi + \frac{1}{3} r/\phi ^2)$
7/2	$\exp(- r/\phi)(1 + r/\phi + \frac{2}{5} r/\phi ^2 + \frac{1}{15} r/\phi ^3)$

$\mathcal{C}(h)$, that only depends on the distance h . It follows directly, that $\text{Cov}[Y(\mathbf{s}), Y(\mathbf{s})] = \mathcal{C}(0)$, and then $\text{Var}[Z(\mathbf{s})] = \sigma_Z^2$, and it is not a function of the spatial location \mathbf{s} .

Isotropy

Another common assumption used to simplify the covariance structure of a spatial process, is the *isotropy*, it consists in assuming that

$$\text{Cov}[Z(\mathbf{s}), Z(\mathbf{s} + h)] \text{ depends only on } \|h\|.$$

This assumption is stronger than stationarity, since it implies that the covariance between observations located at $\|h\|$ units apart is the same, independently of the location and geographical direction (North-South or East-West). In terms of a kriging model in (1.2), implies that:

$$C = \mathcal{C}(r), \text{ for } 1 \leq i, j < n, \text{ and } r = \|\mathbf{s}_i - \mathbf{s}_j\|,$$

where

$$\mathcal{C}(r) = \sigma_Z^2 \rho(r), \quad \sigma_Z^2 = \text{Var}[Z(\mathbf{s})],$$

and \mathcal{C} and ρ are respectively the covariance and correlation functions for the isotropic process Z . The correlation function ρ satisfies that $\rho(0) = 1$, thus, the selection of the correlation function is needed to ensure the positive definiteness of the covariance function C .

Matérn family of covariance functions

Stein (1999) proposed the use of a class of functions based on the spectral decomposition of the covariance function. The common choice is the Matérn class family (Matérn, 1986)

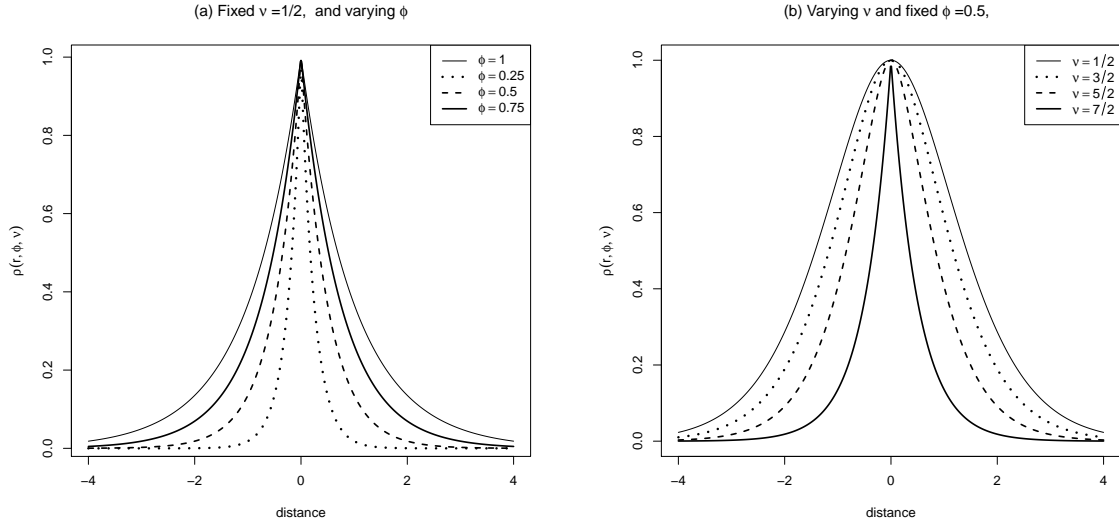


Figure 1.5: Matérn correlation functions, for different values of ν and range parameter ϕ .

given that they allow a great flexibility and they are characterized by the general form:

$$\rho(r; \phi, \nu) = \left(2^{(\nu-1)}\Gamma(\nu)\right)^{-1} \left(\frac{r}{\phi}\right)^{\nu} \mathcal{K}_{\nu}\left(\frac{r}{\phi}\right), \quad (1.5)$$

where Γ is the Gamma function and \mathcal{K}_{ν} is the modified Bessel function of order $\nu > 0$. The expression in (1.5) has no closed form for general ν , however for particular values it is possible to obtain simple form, using that $\nu = m + \frac{1}{2}$, for $m = 0, 1, 2, \dots$, where $0 < \nu < \infty$ (see Table 1.1). In the limit (as $\nu \rightarrow \infty$), the Gaussian correlation functions is obtained. For $\nu = \frac{1}{2}$, the correlation function is exponential. The parameter ϕ is a scale or range parameter that determines how fast the correlation decays when the distance increases. Figure 1.5 illustrates the shape of the Matérn correlation functions for different values of ν and ϕ . The estimation of the Matérn family parameters (σ_Z^2 , σ^2 , ρ and ν) can be done using likelihood based methods (Stein, 1999; Nychka, 2000). The parameter ν is a *smoothness parameter* that allows flexible covariance structures.

The main disadvantages of kriging methods, arise from the strong assumptions needed for consideration. In practice, real data sets (specially those data related to natural of physical phenomena) are rarely stationary or isotropy. Even when non-stationary covariance functions are defined, they are restricted to a few unrealistic situations.

Additionally, these techniques requires the specification of a proper spatial covariance model. In terms of computational efficiency for large data sets, the kriging algorithm has to solve numerous simultaneous equations in order to obtain the best linear

predictor. Recently, [Cressie and Johannesson \(2008\)](#) proposed a more flexible family of non-stationary covariance functions based on a reduced number of basis functions, that are fixed in number and allow for the reduction of the computational cost when the number of spatial points n is large. We follow a similar low-rank models approach using non-parametric models we will discuss later in Section 1.4.

1.2.2 Spatial models for regional data

In regional data, the spatial structure is incorporated in the modelling through the concept of spatial neighbors. This is similar to times series analysis, where an observation at time t is a linear combination of the past observations. The spatial analogous, consists in considering that the observation at location s is a linear combination of their neighbors. In the spatial context, these models are known as spatial autoregressive models, since the autoregression induces a spatial correlation among the regions that are nearby. As a difference to the geostatistical case, the main objective is not to predict values at new locations, but to study the existence of a spatial pattern (i.e. if regions that are near to each other tend to take similar values than regions far from each other).

[Besag \(1974\)](#) proposed to follow the ideas of autoregressive models in times series analysis to the spatial context. In times series, the sequence of random variables Y_1, Y_2, \dots, Y_T , is said to have the *Markov property* if the conditional distribution of Y_{t+1} given Y_1, Y_2, \dots, Y_t is the same as as the conditional distribution of Y_{t+1} given Y_t . This means that the value at time $t + 1$ depends only on the previous value. The extension of this property to the spatial case means that for given spatial process $Y(s)$, the full conditional distribution at each observation s_i is such that:

$$\Pr(Y(s_i)|Y(s_j), j \neq i) = \Pr(Y(s_i)|Y(s_j), j \in \mathcal{N}_i), \quad i = 1, \dots, n. \quad (1.6)$$

In other words, $Y(s_i)$ depends on $Y(s_j)$ if and only if location s_j is in the neighborhood set \mathcal{N}_i of s_i . In the spatial, context $Y(s)$ is defined as a *Markov Random Field* (MRF).

Let us define the spatial process in terms of a Gaussian linear regression model, i.e.

$$Y(s) = \mathbf{X}(s)\boldsymbol{\beta} + \epsilon(s), \quad (1.7)$$

and $\epsilon(s) \sim \mathcal{N}(0, \boldsymbol{\Omega})$. The spatial dependence is modelled through the covariance matrix $\boldsymbol{\Omega}$. We consider the conditional distributions are Gaussian, then the first two conditional

moments are:

$$\mathbb{E}[Y(\mathbf{s}_i)|Y(\mathbf{s}_j), j \neq i] = \mathbf{X}(\mathbf{s}_i) + \sum_{j=1}^n c_{ij}(Y(\mathbf{s}_j) - \mathbf{X}(\mathbf{s}_j)) \quad (1.8)$$

$$\text{Var}[Y(\mathbf{s}_i)|Y(\mathbf{s}_j), j \neq i] = \sigma_i^2, \quad i = 1, \dots, n, \quad (1.9)$$

where c_{ij} are the spatial dependence parameters that are generally specified through the neighbourhood structure, i.e. $c_{ii} = 0$ and $c_{ij} \neq 0$ if $j \in \mathcal{N}_i$ and zero otherwise. To ensure that the covariance matrix is symmetric, it is necessary to impose the constraints: $\sigma_j^2 c_{ij} = \sigma_i^2 c_{ji}$. Then, for the Gaussian case and given (1.8) and (1.9), we have that the joint distribution of Y is defined by:

$$Y(\mathbf{s}) \sim \mathcal{N}(\mathbf{X}(\mathbf{s})\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\text{CAR}}), \quad (1.10)$$

where $\boldsymbol{\Sigma}_{\text{CAR}} = (\mathbf{I} - \mathbf{C})^{-1}\boldsymbol{\Sigma}_c$, and $\mathbf{C} = \{c_{ij}\}$ is an $n \times n$ matrix, and $\boldsymbol{\Sigma}_c = \sigma^2 \mathbf{I}$.

The matrix \mathbf{C} is usually expressed in terms of a parametric function of the spatial connectivity matrix W as defined in (1.1), e.g. $\mathbf{C} = \rho W$, and also $\boldsymbol{\Sigma}_c = \sigma^2 \mathbf{V}_c$, with \mathbf{V}_c known. Then, the CAR covariance matrix can be written as $\boldsymbol{\Sigma}_{\text{CAR}} = \sigma^2 \mathbf{V}_c(\rho)$. Using these simplifications, the estimation of $\boldsymbol{\beta}$, ρ and σ^2 can be done by maximization of the likelihood (see Cressie, 1993, pg. 408).

CAR models are very popular in the study of spatial (regional) patterns of a disease. This area of research is known as *disease mapping*, where the response variable $Y(\mathbf{s}_i)$ are counts of observed number of cases of a disease in county i , for $i = 1, \dots, n$. We will discuss these models in Chapter 3.

1.2.3 Point patterns models

The analysis of spatial point patterns is classified in two groups (Haggett, 1977): (i) the methods based on *distances*, that use the information of the spatial locations to characterize the spatial pattern (usually with the mean distance to the nearest neighboring point), and (ii) methods based on *areas*, that divides the domain \mathcal{D} in smaller sub-regions of equal size (quadrats) and study the spatial pattern counting the number of events per unit area within a quadrat. The process of counting the number of events is used to represent the frequency distribution of the observed numbers of points per quadrat.

A spatial point patterns is denominated as *complete random pattern* if the average number of events at area unit is homogeneous across the domain \mathcal{D} . A complete spatial randomness implies that an event is equally probable to occur in any location of the area of study regardless of the locations of other events. Then, the number of events in two no overlapping areas A_1 and A_2 are independent and follow a Poisson distribution. The

interest of spatial point pattern analysis is to study if events are uniformly distributed or not and determine if there exists clusters of some regular pattern in the locations of events.

The analysis of spatial point patterns consider the *intensity function*, $\lambda(\mathbf{s})$ as the average density points or the expected number of points per unit area. A spatial Poisson process is *heterogeneous*, if the intensity function $\lambda(\mathbf{s})$ varies spatially, and it has two properties:

- (i) Let $N(A)$ be the number of events in an area $A \subset \mathcal{D}$, then

$$N(A) \sim \text{Poisson}(\lambda(A)),$$

where $\lambda(\mathbf{s})$ is the intensity function at location \mathbf{s} , where $(0 < \lambda(\mathbf{s}) < \infty)$, and

$$\lambda(A) = \int_A \lambda(\mathbf{s}) d\mathbf{s}.$$

- (ii) If A_1 and A_2 are two disjoint or no over-lapping areas of \mathcal{D} , then $N(A_1)$ and $N(A_2)$ are independent.

A spatial point pattern is characterized by two properties:

- (i) *First-order property of the intensity function*, that describes the way in which the mean of the process varies across space, and measures the number of events per unit area. Formally, Diggle (1983) defines it as the average number of events per unit area at location \mathbf{s} :

$$\lambda(\mathbf{s}) = \lim_{|\delta\mathbf{s}| \rightarrow 0} \left\{ \frac{\mathbb{E}[N(\delta\mathbf{s})]}{|\delta\mathbf{s}|} \right\}, \quad (1.11)$$

where $\delta\mathbf{s}$ is a small region around \mathbf{s} , $|\delta\mathbf{s}|$ is the area of this region, and $N(\delta\mathbf{s})$ is the number of events in the small region $\delta\mathbf{s}$.

- (ii) *The Second-order property*, describes the spatial dependence between the observations in pairs of subregions within the domain \mathcal{D} . It is defined as:

$$\gamma(\mathbf{s}_i, \mathbf{s}_j) = \lim_{|\delta\mathbf{s}_i|, |\delta\mathbf{s}_j| \rightarrow 0} \left\{ \frac{\mathbb{E}[N(\delta\mathbf{s}_i)N(\delta\mathbf{s}_j)]}{|\delta\mathbf{s}_i||\delta\mathbf{s}_j|} \right\}. \quad (1.12)$$

These properties, allow us to define the concepts of *stationarity* and *isotropy*, as seen in Section 1.1.1, for spatial point patterns. A spatial point pattern is stationary, when it is invariant to translation, i.e. $\lambda(\mathbf{s}) = \lambda$ (or equivalently if it is homogeneous). Formally,

the second-order intensity depends only on event location differences, i.e. $\gamma(\mathbf{s}_i, \mathbf{s}_j) = \gamma(\mathbf{s}_i - \mathbf{s}_j)$. If the spatial process is isotropic, the second-order intensity function depends only on distance, i.e. if $\gamma(\mathbf{s}_i, \mathbf{s}_j) = \gamma(\|\mathbf{s}_i - \mathbf{s}_j\|) = \gamma(h)$.

Kernel estimation of spatial point patterns

The study of the intensity function has been mostly developed in non-parametric smoothing techniques. A generalization of the area-based methods is the estimation of the intensity function using *kernels*. Kernel methods are mathematical tools used in many areas of statistics, to obtain smooth estimates of probability density functions (univariate and multivariate) from an observed sample of observations (Silverman, 1986; Wand and Jones, 1994).

In the context of spatial point pattern analysis, the estimation of the intensity function is similar to estimate a bivariate probability density function. The density estimation produces an estimate of the probability of observing an event at location \mathbf{s} and integrates to one over the sub-region A .

The intensity function $\lambda(\mathbf{s})$, at \mathbf{s} estimated at \mathbf{s}_0 , is

$$\hat{\lambda}_\tau(\mathbf{s}_0) = \sum_{i=1}^n \frac{1}{\tau^2} \mathcal{K}\left(\frac{\mathbf{s}_i - \mathbf{s}_0}{\tau}\right), \quad (1.13)$$

where $\mathcal{K}(\cdot)$ is the standardized kernel weighting function (centered at \mathbf{s} and with unit volume). The value $\tau > 0$ is the *bandwidth* or neighborhood and measures the distance of an observed event \mathbf{s}_i that lies within the region of interest. Figure 1.6 illustrates the kernel estimation of a point pattern. The selection of the kernel function is less relevant than the choice of the bandwidth h . Several authors have discussed the appropriate choice of the kernel and selection criteria for h (see Wand and Jones, 1994; Ruppert et al., 1995; Hall et al., 1995), for spatial context see Diggle (1981, 1983, 1985).

However, kernels suffer from some drawbacks as for example: *edge effects* at boundaries, that can lead to biased estimates close to the boundary of \mathcal{D} , since events close to the boundary has no neighboring events outside \mathcal{D} . Solutions to this effect is to include edge-correction terms (see Diggle, 1981; Zheng et al., 2004).

Figure 1.7, illustrates the kernel estimation of the intensity function for different values of the bandwidth. We chose a Gaussian kernel (i.e. $\mathcal{K}(x) = (2\pi)^{-0.5} \exp\{-x^2/2\}$), for the maples trees of Lansing woods data showed in Section 1.1.3.

Another alternative to estimate the intensity function parametrically by the use of likelihood methods. Let us suppose a realization of n independent events of an hetero-

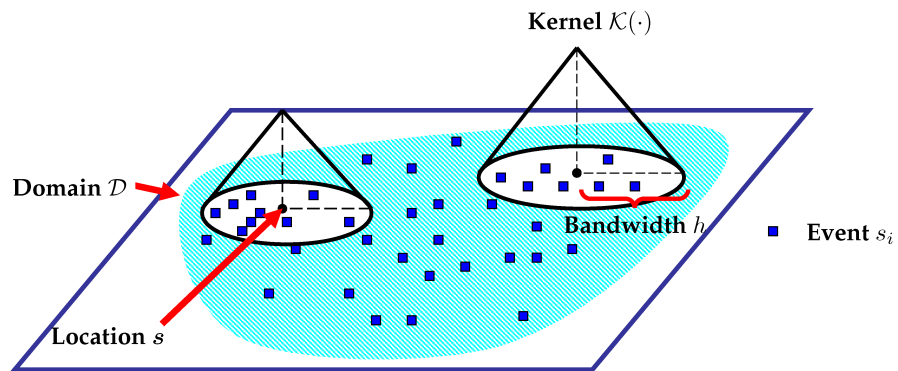
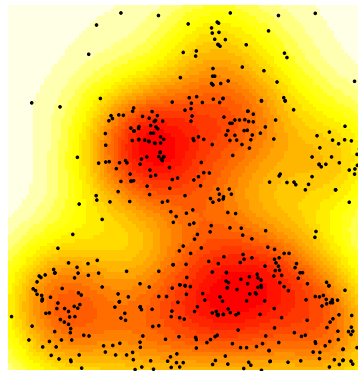
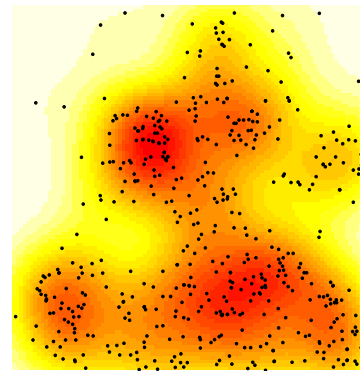
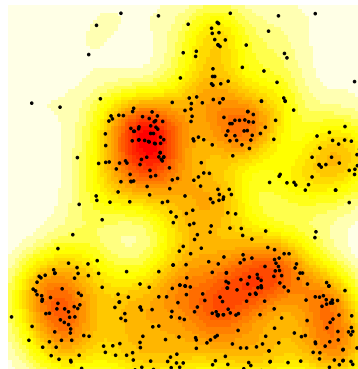
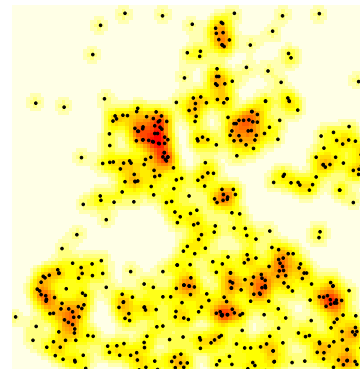


Figure 1.6: Kernel estimation in spatial point patterns

(a) $h = 0.1$ (b) $h = 0.08$ (c) $h = 0.06$ (d) $h = 0.02$ Figure 1.7: Estimated kernel density estimates of maple trees from the Lansing data set with Gaussian kernel and $h = \{0.1, 0.08, 0.06, 0.02\}$.

geneous Poisson process with intensity function $\lambda(\mathbf{s})$, the log-likelihood is:

$$\mathcal{L}(\lambda) = \sum_{i=1}^n \log \lambda(\mathbf{s}_i) - \int_A \lambda(\mathbf{s}) d\mathbf{s}, \quad (1.14)$$

where $\int_A \lambda(\mathbf{s}) d\mathbf{s}$ is the expected number of cases of the heterogeneous Poisson process with intensity $\lambda(\mathbf{s})$ in region A . Diggle (2003) suggests the use of a log-linear model of the form:

$$\log \lambda(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta},$$

where \mathbf{X} is the $n \times p$ matrix of covariates at location \mathbf{s} , and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$, the vector of parameters. The estimates of the parameters can be obtained by maximization of the log-likelihood in (1.14), for practical details see Baddeley and Turner (2000).

1.2.4 Hierarchical spatial models

The popularity of spatial models has increased in the last decade, since they are easy to implement in the context of *Gibbs sampling* and *Markov Chain Monte Carlo* (MCMC) framework. Hierarchical modelling is based on the fact that the joint distribution of a set of random variables can be decomposed into a series of conditional models. These models allow to incorporate various sources of uncertainty to accommodate complex relationships between the data and the random process.

A basic hierarchical model is represented in three stages:

- **Stage 1.** Data model: $[Data|Process, data\ parameters]$. This stage specifies the distribution of the data given the process of interest and parameters that describe the data model.
- **Stage 2.** Process model: $[Process|Process\ parameters]$. The second stage, describes the process conditional on other process parameters, and
- **Stage 3.** Parameter model: $[Data\ and\ process\ parameters]$, is to model the uncertainty in the parameters, from both stages 1 and 2.

Bayesian methods are the natural way to consider this hierarchical setting. Using the Bayes theorem, we can obtain the *posterior distribution* (i.e. the joint distribution of the process and parameters given the data), which is proportional to the data model (i.e. the likelihood), times the *prior distribution*, i.e.:

$$[Process, parameters|Data] \propto [Data|Process, parameters] \times [Process|parameters][parameters]. \quad (1.15)$$

The estimation of the posterior is therefore done by MCMC through iterative sampling. The software for Bayesian computation WinBUGS¹ implements these models using a high-level programming language.

A hierarchical spatial model for a Gaussian spatial process, can be written as a general linear regression model as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon}, \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \text{ and } \mathbf{u} \sim \mathcal{N}(0, \mathbf{R}), \quad (1.16)$$

where $\mathbf{y} = (y(\mathbf{s}_1), \dots, y(\mathbf{s}_n))'$, represents the spatial process at location i , for $i = 1, \dots, n$, and \mathbf{u} is a vector of spatial random effects with covariance \mathbf{R} , and error term $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \tau^2 \mathbf{I})$. The covariance matrix \mathbf{R} , can be defined as $\mathbf{R} = \sigma^2 \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a $n \times n$ correlation matrix that takes into account for the spatial correlation. For spatially continuous data, we can define $\Lambda_{ij} = \rho(r; \phi, \nu)$, with $r = \mathbf{s}_i - \mathbf{s}_j$, and $\rho(r; \phi, \nu)$ is a valid Matérn isotropic correlation function as defined in (1.5). Let us consider $\boldsymbol{\theta}$ as the vector of all model parameters: $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \phi, \nu)'$. Then, the hierarchical stages are:

$$\mathbf{y} | \mathbf{u}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \sigma^2 \mathbf{I}) \quad (1.17)$$

$$\mathbf{u} | \sigma^2, \phi, \nu \sim \mathcal{N}(0, \sigma^2 \boldsymbol{\Lambda}). \quad (1.18)$$

The model specification is completed by adding priors for $\boldsymbol{\beta}$ and τ^2 , as well as for the hyperparameters σ^2 , ϕ and ν .

For regional data, the formulation of hierarchical CAR model, requires the specification of the covariance matrix \mathbf{R} in terms of the neighborhood connectivity matrix as defined in Section 1.2.2 (see Banerjee et al., 2004, for more details).

This general scheme can be used for non-Gaussian responses in a generalized linear framework, extending the methodology to the exponential family. Thus, the process is written as: $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon}$, where $\boldsymbol{\eta} = g(\boldsymbol{\mu})$ is the linear predictor and $g(\cdot)$ is a known link function of the mean $\boldsymbol{\mu}$ of the spatial process. Then, for spatial point patterns we can consider data from a Poisson distribution with:

$$\mathbf{y} | \boldsymbol{\lambda} \sim \text{Poisson}(\boldsymbol{\lambda}),$$

where $\boldsymbol{\lambda} = (\lambda(\mathbf{s}_1), \lambda(\mathbf{s}_2), \dots, \lambda(\mathbf{s}_n))'$ is the vector of unknown spatial Poisson intensity function. The Poisson intensity process can then be modelled in the process stage (stage 2) using the information of the covariates.

¹<http://www.mrc-bsu.cam.ac.uk/bugs/>

1.3 Spatio-temporal modelling

In previous sections, we introduced the main concepts and modelling approaches in the spatial statistics literature. In this Section, we extend the scope to the temporal domain. In recent years, there has been an enormous growth of data with spatial structure that are temporally indexed. This type of data arise in many contexts such as, meteorology, environmental sciences, epidemiology or demography, among others. This wide variety of settings has generated a considerable interest in the development of spatio-temporal models. However, the complexity of the models needed and the size of the data sets has made this a challenging task.

The classification of spatial data showed in [Section 1.1](#), can be extended to the spatio-temporal case. A spatio-temporal process, $Y(s, t)$, is defined as:

$$\{Y(s, t), \text{ where } s \in \mathcal{D}, t \in T\},$$

with $s \in \mathcal{D} \subset \mathbb{R}^2$, $T \subset \mathbb{R}$. Then, the realization of the spatio-temporal process $Y(s, t)$ has been collected in s locations and over $t = 1, \dots, T$ time points.

Figure 1.8a presents the locations of the monitoring stations, and Figure 1.8b the seasonal pattern in ozone levels in four different countries (Spain, Sweden, Austria and UK). The plots show that the stations cover a large area where spatial trends are likely to appear (mostly due to climate conditions), and a clear seasonal pattern is present along the years. Now the interest lies not only in studying the spatial surfaces of ozone levels in a particular time, but also how this surface changes over time. The measurements are collected hourly, daily, etc..., over several years, and therefore, appropriate models should be able to study both spatial and temporal trends.

From a methodological point of view, the incorporation of the temporal component involves a significative increase in model complexity. We need to study not only the spatial dependence structure, but also the temporal and space-time interaction dependence structure. For example, in geostatistical data, the kriging methods in [Section 1.2.1](#) need to specify a space-time covariance structure. In general, given two spatio-temporal processes, let us say $Y(s_1, t_1)$ and $Y(s_2, t_2)$, will depend on additional assumptions as stationarity and separability. It is common to assume under certain conditions, that the covariance structure of a spatio-temporal process is *separable*, such that:

$$\text{Cov}[Y(s_1, t_1), Y(s_2, t_2)] = \text{Cov}_S(s_1, s_2) \times \text{Cov}_T(t_1, t_2) \quad (1.19)$$

for each space-time coordinate (s_1, t_1) and (s_2, t_2) in $\mathbb{R}^d \times \mathbb{R}$, and where Cov_S and Cov_T , are a purely spatial and temporal covariances. The main advantage of the assumption

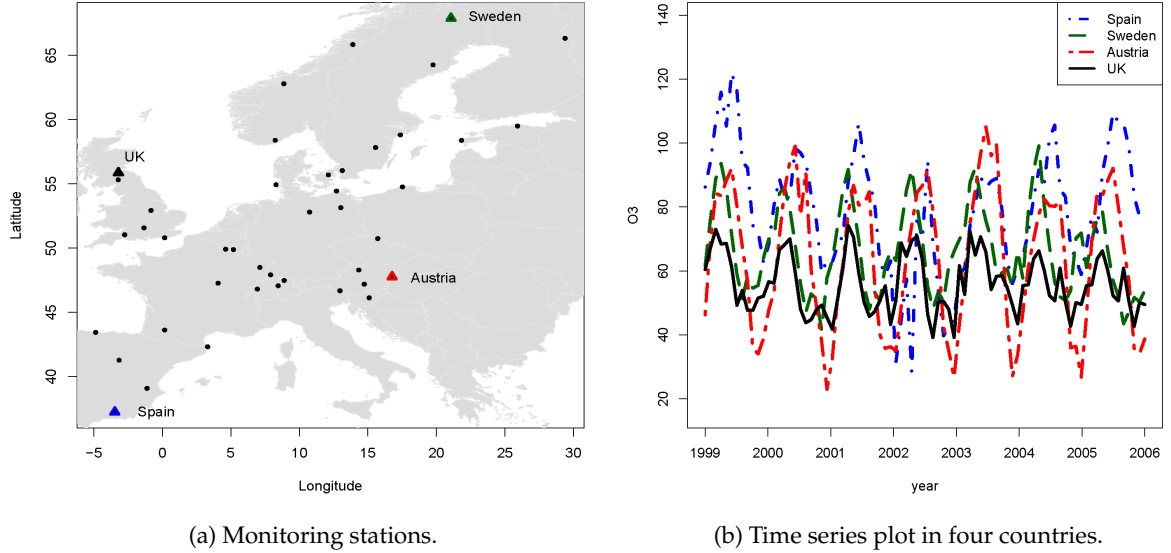


Figure 1.8: (a) sample of 43 monitoring stations over Europe. (b) O_3 levels in four selected countries.

of separability is computational, and it also offers an attractive interpretation (Mardia and Goodall, 1993; Goodall and Mardia, 1994). However, this covariance assumes that both structures can be modelled separately, and then does not consider the space-time interaction. Also, as noted by Stein (2005), in general separable covariance functions imply that small changes in the locations of the observed values may lead to large changes in the correlations between certain linear combinations of observations.

Cressie and Huang (1999) developed some class of non-separable stationary covariance functions to model the space-time interaction. However, their approach is restricted to a small class of valid functions for which a Fourier transform integral is known. Gneiting (2002) overcame these difficulties and provided a more general class of valid space-time covariance models. Fuentes et al. (2008) used the spectral decomposition of a spatio-temporal process to develop a flexible class of parametric space-time covariance models. Huang and Cressie (1996) developed a dynamic model using the Kalman filter for separable covariance structures, and Mardia et al. (1998) implemented a reduced dimension model (“kriged Kalman filter”) for modelling non-separable process that can be applied to large data sets.

In the Bayesian framework, spatio-temporal processes are modelled by the implementation of the hierarchical Bayesian methodology. Now, the stages of the hierarchical spatio-temporal model incorporates the uncertainty in the observations, in the specifi-

cation of the spatio-temporal process, and in the knowledge of the parameters that describe the space and time dependence. However, the use of MCMC algorithms for this very high-dimensional models lead to computationally intensive and complex models. The implementation of the MCMC to spatio-temporal problems have been discussed by several authors (Waller et al., 1997; Wikle et al., 1998; Gössl et al., 2001; Banerjee et al., 2004), and different variations of the MCMC scheme have been proposed. They present non-separable hierarchical models based on Markov random fields in which both dependence structures are incorporated through the prior. In these models the interaction is modelled by Kronecker products of precision matrices. However, these approaches assume isotropic processes, which is unrealistic in many cases. The use of the hierarchical Bayesian modelling approach is a challenging task, in spatio-temporal data due to the high-dimensionality of the problem and given that the MCMC implementation has slow convergence and long computing times.

1.4 The smoothing approach

The main focus in the previous sections was to introduce the typology of spatial data and most of the common models applied for them. From a statistician point of view, the study of variables is done by the so-called regression techniques, where a response variable, y_i , $i = 1, \dots, n$, is explained by a multiple predictor variables $x = (x_1, \dots, x_k)'$. In linear regression, the mean surface is a plane in the sample space, that in most of the cases are too simplistic because of the non-linearity in the data. The aim of non-parametric regression techniques is to extend the linear regression to more flexible and *smooth* forms for the mean surface, whose exact form is not pre-specified but chosen from a flexible family of fitting procedures. The non-parametric regression model consists in:

$$y = f(x) + \epsilon, \quad (1.20)$$

where $f(\cdot)$ is a smooth, unknown continuous function, and ϵ is the error *i.i.d.* term, such that $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The literature on non-parametric regression techniques or simply smoothing techniques is very wide (see Hastie and Tibshirani, 1990; Simonoff, 1996; Bowman and Azzalini, 1997, for a detailed review) or *lowess* methodology (Cleveland and Devlin, 1988).

Among other smoothing techniques, *spline* fitting is a popular method of interpolation (see Wahba, 1990; Green and Silverman, 1994). Several authors as Laslett (1994); Mardia et al. (1996); Nychka (2000) have discussed the connection of kriging and splines and presented comparisons of the performance of both methods. In the spatial case, the general model formulation of the smooth function $f(\cdot)$ in (1.20) can be considered from

the smoothing approach as a bivariate function of the spatial covariates $\mathbf{s} = (x_1, x_2)$, where x_1 and x_2 represents in general the geographical coordinates (longitude and latitude).

Splines are piecewise polynomial functions constrained to join at certain points called *knots*, that are evenly spaced through the range of observed values of x . The basic elements to be considered for spline fitting are: (i) degree of the polynomial; (ii) the number of knots; and (iii) the location of knots. The popular choice is the cubic spline.

Let us consider the set of points x_1^*, \dots, x_n^* in the interval $[a, b]$, such that:

$$a < x_1^* < x_2^* < \dots < x_n^* < b.$$

The function $f(\cdot)$, defined in the interval $[a, b]$ is a *cubic spline*, if satisfies the next conditions:

- (i) For each of the intervals $(a, x_1^*), (x_1^*, x_2^*), (x_2^*, x_3^*), \dots, (x_n^*, b)$, f is a cubic polynomial,
- (ii) each of the pieces of the polynomial join at points x_i^* , in such a way that the function f and its first and second order derivatives are continuous at each point x_i^* and therefore within the interval $[a, b]$. The points x_i^* are the *knots*.

A *natural spline*, includes additional constraints, such that $f''(a) = f''(b) = 0$, i.e. that the function is linear beyond the boundaries.

Non-parametric regression with splines, are commonly known as *Smoothing splines*. Consider the regression problem in (1.20), the smoothing splines are the solution to the problem of minimizing the *residual sum of squares*:

$$\text{RSS}(f, \lambda) = \sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \int f''(x)^2 dx, \quad (1.21)$$

where the first term measures the closeness of the fit to the data, and the second term penalizes the wiggleness of the function f , and λ is a smoothing parameter that establishes the trade-off between data fitting and smoothness of f . The smoothing parameter $\lambda \in (0, \infty)$ and if $\lambda = 0$, f can be any function that interpolates the data, and if ∞ , the second derivative is constrained to 0, and the fit corresponds to the least squares solution (a straight line). Smoothing splines are natural cubic splines with all knots at unique values of x . The smoothing spline estimator in the sense that for each unique x , there are basis functions, $b_i(x)$, for $i = 1, \dots, n$, such that,

$$f_\lambda(x) = \sum_{i=1}^n b(x_i) \beta_i.$$

Then the RSS can be written as:

$$RSS(\beta, \lambda) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda \int f''(x_i)^2 dx,$$

where $\mathbf{X} = [1, b_1(x), \dots, b_n(x)]$, and $\beta = (\beta_1, \beta_2, \dots, \beta_n)'$ is the vector of regression coefficients. Given that f is linear in the parameters β_i , the penalty can be written as a quadratic form in β , as:

$$\int f''(x_i)^2 dx = \beta' \Omega \beta,$$

where Ω is a matrix of known coefficients. Then, given the smoothing parameter λ , the solution of (1.21) is:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X} + \lambda\Omega)^{-1} \mathbf{X}'\mathbf{y}.$$

Then, the smoothing spline estimator is now a problem of estimating the smoothing parameter λ . We will see this issue in Section 2.1.3 of Chapter 2.

Another type of spline based smoothing technique are the *Thin plate splines* (Duchon, 1976). Thin plate splines are the natural analog of the cubic spline in several dimensions and the penalty term in this case is:

$$J_m(f) = \int_{\mathbb{R}^d} \sum \frac{m!}{\alpha_1! \dots \alpha_d!} \left(\frac{\partial^m f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \right)^2 dx_1 \dots dx_d, \quad (1.22)$$

where expression (1.22), is computed over all non-null α 's, such that $\sum \alpha_1 + \dots + \alpha_d = m$, and $2m > d$. For the bivariate case, $J_2(f) = \int f''(x)^2$, that yields the penalty for a cubic spline.

The main disadvantage of Smoothing splines and Thin plate splines is computational, since it requires to estimate as many unknown parameters as data. Recently, Wood (2003) proposed the *Thin plate regression splines* as a computationally efficient version of Thin plate splines. Another issue in these type of spline models for regression is that considering an unique smoothing parameter, λ , we are considering an *isotropic* smoothing.

The popularity of splines in regression problems has exploded in recent years, due to the use of low-rank basis functions for smoothing (Hastie and Tibshirani, 1990; Hastie, 1996). The idea is to use a reduced basis for the regression, such that the regression parameters are lower than the number of data. Recently, low-rank models have been applied in the spatial and spatio-temporal context. For instance, the geosadditive models (Kammann and Wand, 2003) or low-rank kriging models to deal with large data sets (Cressie and Johannesson, 2008). We will consider the low-rank approach in this thesis,

using penalized splines regression models (Eilers and Marx, 1996). We consider the representation of the low-rank models using a mixed model formulation. We study in further details this methodology in Chapter 2.

The use of smoothing techniques for spatial data can also be considered for regional data. It can be considered that the summary data of each region have been observed at the centroid of the region, and the distances between centroids are used to construct the spatial covariance structure as in geostatistical models (see Cressie and Chan, 1989; Cressie, 1993; Berke, 2004; Wall, 2004, for a further discussion). For spatial point patterns, the kernel methods are essentially non-parametric, recent works by Bell and Grunwald (2004) included this analysis in the context of generalized linear mixed models. In Section 3.3 of Chapter 3, we will present examples of spatial data smoothing using the Penalized splines methodology in Chapter 2 for each of the types of spatial data (geostatistical, regional and point patterns).

Most of the common approaches in spatio-temporal data smoothing are considered in the additive models framework. They extend the geoadditive models proposed by Kammann and Wand (2003), or assume a smooth function to model non-linear time effects (MacNab and Dean, 2001; Fahrmeir et al., 2004; Kneib and Fahrmeir, 2006). This formulation implies that the response variable y is modelled as the sum of spatial and temporal effects of the form:

$$\mathbb{E}[y] = f(\text{space}) + f(\text{time}) .$$

This additive model, does not account for the space-time interaction effect, and therefore, can not reflect important features in the data. In general, this assumption implies a spatio-temporal correlation structure given by separable covariance terms for a spatial and temporal components respectively. This approach is computationally very attractive but results too simplistic in real situations. In a very recent work Bowman et al. (2009), consider spatio-temporal models within the additive framework for sulphur dioxide (SO_2) pollution over Europe. The space-time structure is constructed by the residuals of the additive model and incorporated in a general spatio-temporal formulation, the interaction terms involving time and seasonal effects are also considered in their study.

As the number of data observations is large (measured in n locations along t time points), the computational issues in spatio-temporal data analysis becomes crucial, and low-rank models are an important modelling tool for this type of data sets. Chapter 4 is devoted to the development of multidimensional smooth low-rank models based on the decomposition of the functions in terms of additive components with interactions.

We apply these models to the spatio-temporal setting. We propose more realistic models which allow for the consideration of the three-dimensional interaction effect. We describe non-separable models for smoothing across spatial and temporal dimension simultaneously, which explicitly consider the interaction between space and time, and may easily be set into computationally efficient methods. Models with functional form that includes the space-time interaction as:

$$f(\text{space}, \text{time}). \tag{1.23}$$

We consider spatially anisotropic models, allowing for different amount of smoothing for spatial coordinates, and also for temporal dimension, and extend model (1.23) to explicitly consider different smooth additive terms for space and time, and space-time interaction.

Law 2: Organize.
"Organization makes a system
of many appear fewer". John Maeda

Chapter 2

Smoothing mixed models

This Chapter introduces the approach of smoothing with penalized splines (Eilers and Marx, 1996) as a mixed model. We begin with the main aspects of the methodology from the univariate Gaussian case, and its extensions to more general responses in the generalized linear models (GLM) framework. In Section 2.2, we present more details of the methodology and its reparameterization as a mixed model. This allow us to unify the mixed model approach through the rest of the chapters. We called this approach: *smoothing mixed models*. Section 2.3 extends the methodology to the multidimensional case.

2.1 Penalized splines: an introduction

Penalized regression splines (Eilers and Marx, 1996), or commonly known as P -splines, have become a flexible and powerful smoothing tool in different areas of research (see Ruppert et al., 2009, for a detailed review). For simplicity, let us suppose the case of a univariate Gaussian data, with response variable y and regressor x . The smooth model is of the form:

$$y = f(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \quad (2.1)$$

where $f(\cdot)$ is an unknown function, and ϵ are independent and identically distributed (*i.i.d*) errors with variance σ^2 . The function f is assumed to be smooth, and it is estimated from the data points (x_i, y_i) , for $i = 1, \dots, n$.

Eilers and Marx (1996) proposed a simple idea based on two main aspects:

- (i) a regression basis,
- (ii) modify the likelihood function by adding a penalty term over adjacent regression coefficients to control the smoothness of the fit.

Let us consider the model in (2.1) in matrix form:

$$f(x) = B\theta, \quad (2.2)$$

where B is a regression basis constructed from the covariate x , such that, $B = B(x)$, and θ , is the vector of regression coefficients. The coefficients θ , can be obtained solving the least squares problem by minimizing the sum of squares:

$$S = (y - B\theta)'(y - B\theta),$$

and obtain the explicit solution:

$$\hat{\theta} = (B'B)^{-1}B'y. \quad (2.3)$$

However, the solution in (2.3) requires the optimal selection of the number of basis functions, and their location to achieve a smooth fit. Also, this solution tends to overfit the data as more basis functions are used. In order to overcome the problem of overfitting, several authors in the spline literature (see for example [O'Sullivan, 1986](#); [Eubank, 1988](#); [Green and Silverman, 1994](#); [Wahba, 1990](#)) proposed the use of a penalty to control the smoothness of the fit instead of using complex algorithms to determine the optimal number and locations of the knots.

[Eilers and Marx \(1996\)](#) simplified the approach, and instead of using a penalty on the second order derivative, they used a difference of the adjacent coefficients and minimize the *penalized sum of squares*:

$$S_p = (y - B\theta)'(y - B\theta) + \theta'P\theta. \quad (2.4)$$

The term P is the *penalty* that forces the coefficients to vary smoothly, and consequently obtain a smoothed curve. As in any other smoothing technique, we require a *smoothing parameter*, λ , in order to control the amount of smoothness.

Therefore, for a given value of λ , the solution of the penalized sum of squares (2.4), is:

$$\hat{\theta} = (B'B + P)^{-1}B'y. \quad (2.5)$$

Once we have presented a brief introduction of the methodology, we proceed in the next Section to detail the basis and penalties we will use through the rest of the chapters.

2.1.1 Bases and penalties

There are several alternatives for the choice of the regression basis B in (2.2). We follow the work by Eilers and Marx (1996), and use their original proposal of B -splines basis. Other authors as Ruppert et al. (2003) use the truncated power functions, although very simple, they can lead to numerical instability due to poor numerical condition when solving the system of equations in (2.5). In contrast, B -splines are numerically superior and have better properties and extensions (see Ruppert et al., 2003; Eilers and Marx, 2004, for discussion). Usually the number of columns of the basis B is lower than the number of data points, that is why this type of techniques are known as *low-rank smoothers* (Hastie and Tibshirani, 1990). Therefore, an important element of the approach is the choice of a basis function.

B -splines basis

B -splines are very popular among statisticians due to their flexibility and easy computation. B -splines are smoothing splines based on B -spline basis functions. For a more mathematically rigorous explanation and algorithms about B -splines see Dierckx (1993) and de Boor (1978). In summary, B -splines consist of polynomial pieces connected by a set of knots in a particular way. The general properties of a B -spline of order p are:

- it consists of $p + 1$ polynomial pieces, each of degree p ;
- the polynomial pieces join at p inner knots;
- at joining points, derivatives up to order $p - 1$ are continuous;
- the B -spline is positive on a domain spanned by $p + 2$ knots; everywhere else its zero;
- except at the boundaries, it overlaps with $2p$ polynomial pieces of its neighbors;
- at given x , $p + 1$ B -splines are nonzero.

The knots divide the interval of x , over which basis functions are calculated, such that, $x_{\min} = k_1 < k_2 < \dots < k_{m-1} < k_m = x_{\max}$, each interval will be covered by $p + 1$ B -splines of degree p . The total number of knots for the construction of the B -splines will be $m + 2p + 1$, and the number of B -splines in the regression basis, i.e. the number of columns of B , is $c = m + p$.

Although it is possible to choose the locations of the knots. In most of the situations, the suggestion is to use a moderately large number of equally-spaced knots (between

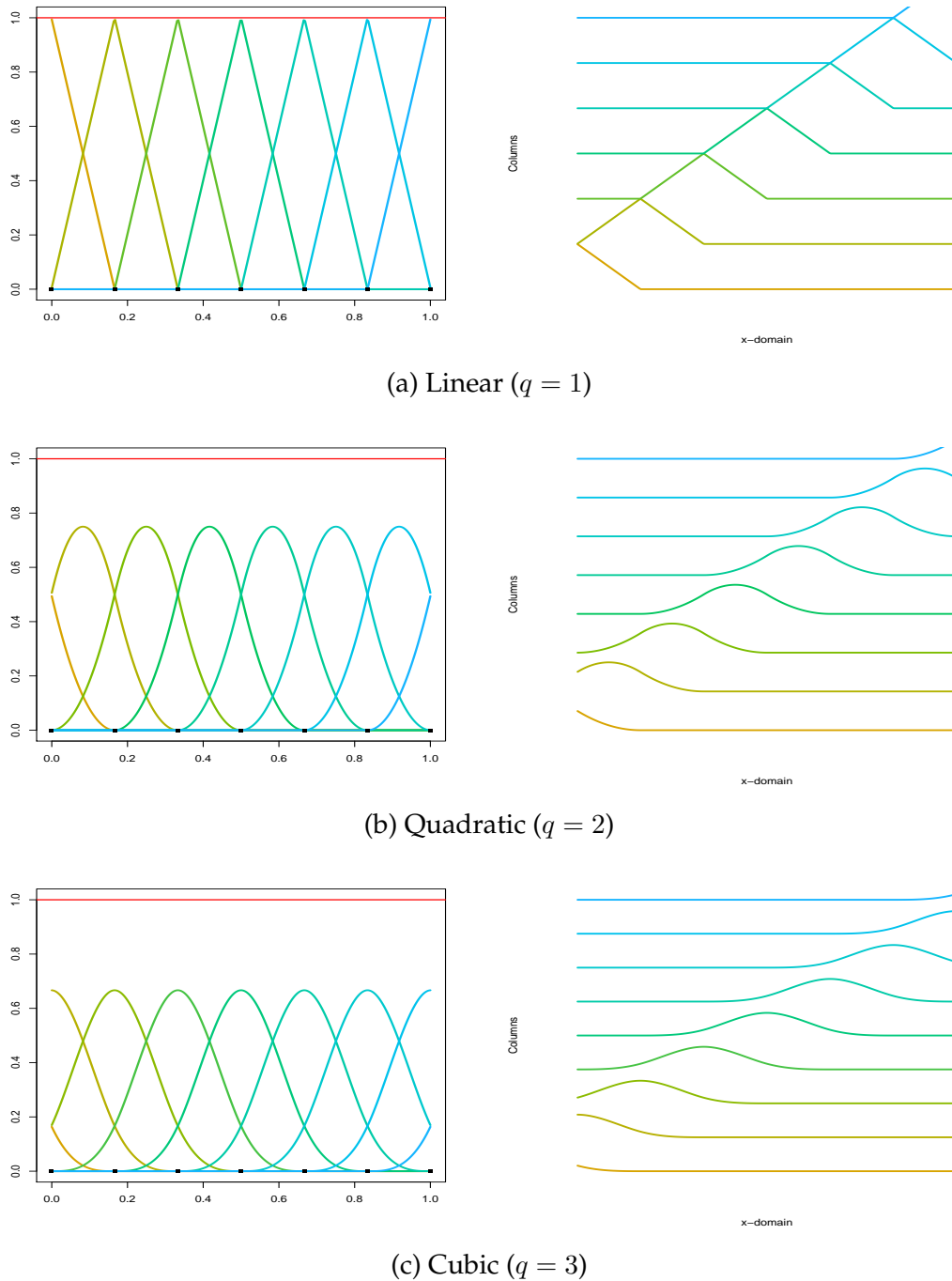


Figure 2.1: B -spline regression bases of different orders of degree p and $m = 6$ and equally-spaced knots

20 and 40). The usual rule is to select:

$$\text{Number of knots} = \min\{1/4 \times \text{unique values of } \mathbf{x}, 40\}.$$

More details about selection of knots in P -splines were studied by [Ruppert \(2002\)](#) and [Ruppert and Carroll \(2000\)](#). Therefore, B is a regression basis formed by several basis functions with the same shape but shifted through the horizontal axis according to the distance between the knots. This property is also true at the boundaries, in contrast with natural splines or kernel-based smoothers. [Figure 2.1](#) shows examples of B regression matrices with $m = 6$ intervals and different degrees for the B -splines. The right panels of [Figure 2.1](#) represent each of the columns of B against the x -domain.

Penalties

Following the approach by [Eilers and Marx \(1996\)](#), the penalty term P , is considered to be discrete, it consists of a difference penalty on the coefficients of the B -spline functions. Therefore, the penalty P in (2.4) is a $c \times c$ matrix of the form $P = \lambda(\Delta^q)' \Delta^q$, where Δ^q is the difference operator of order q . For the vector of regression coefficients θ , the difference operator is defined recursively by:

$$\begin{aligned}\Delta^1 \theta_i &= \theta_i - \theta_{i-1}, \\ \Delta^2 \theta_i &= \Delta^1(\Delta^1 \theta_i) = \theta_i - 2\theta_{i-1} + \theta_{i-2} \\ &\vdots \\ \Delta^q \theta_i &= \Delta^1(\Delta^{q-1} \theta_i).\end{aligned}$$

The order of the penalty q , controls the changes between adjacent coefficients. A first order difference ($q = 1$), penalizes jumps between successive coefficients and a second order difference penalizes deviations from the linear trend (i.e. from $2\theta_{i-1} - \theta_{i-2}$). Therefore, the P -spline fit has an interesting property: a strong smoothing (large values of the smoothing parameter λ) leads to a polynomial of degree $p - 1$, and consequently independent to the degree q of the spline basis (in contrast to the truncated power basis).

In matrix form, we define the matrix D_q as the q^{th} order difference of the vector of regression coefficients θ , i.e. for first and second order differences and $c = 5$, we have:

$$D_1 = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{pmatrix}.$$

Then, the penalty can be written as $P = \lambda D_q' D_q$, where λ controls the amount of the smoothing. The usual choice for the penalty is a second order, $q = 2$, in that case,

we have that the penalty is equivalent to

$$(\theta_1 + 2\theta_2 + \theta_3)^2 + \dots + (\theta_{c-2} + 2\theta_{c-1} + \theta_c)^2 = \theta' D' D \theta. \quad (2.6)$$

Note that, other orders might be more appropriate in some cases. Figure 2.2 illustrates the performance of the P -spline methodology. We simulated $n = 100$, (x_i, y_i) points, from the function $f(x_i) = 1.2 + \sin(5x_i) + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, 0.2)$ and $x_i \sim \text{Unif}[0, 1]$. Figure 2.2 (a) shows the P -spline fit without penalty (i.e. $\lambda = 0$), corresponding to a simple B -spline regression. Figure 2.2 (b) shows the P -spline fit with a penalty (with λ fixed to 10). In both figures, we used a cubic spline for the B -spline basis ($p = 3$), with $m = 20$ knots and a second order penalty ($q = 2$). In both figures we also represent the B -splines bases multiplied by the vector of coefficients θ (represented in circles).

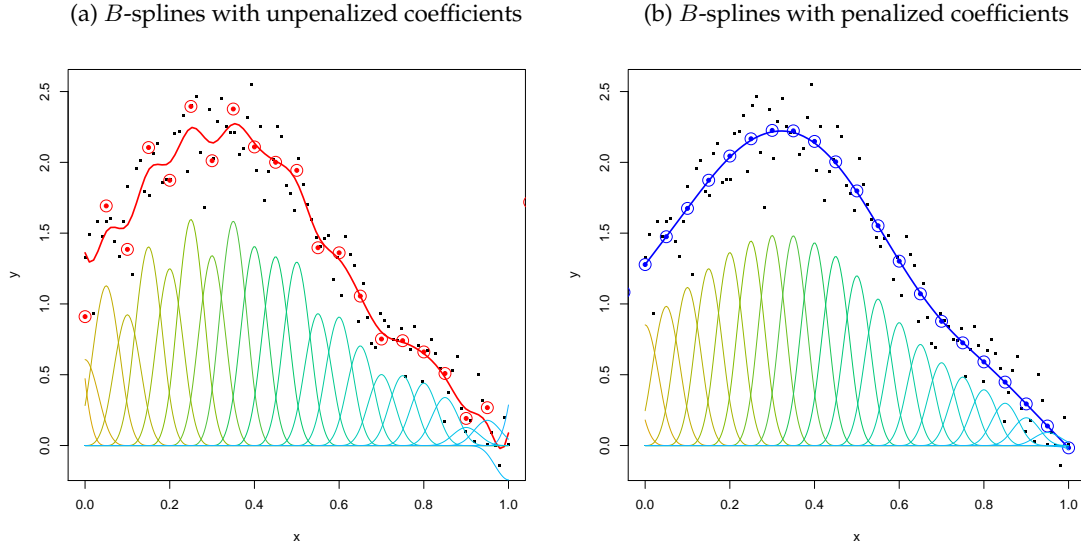


Figure 2.2: (a) fitted curve with unpenalized coefficients (red circles). Bottom: fitted curve with penalized coefficients (blue circles).

It is obvious that the shape of the fitted curve is influenced by the value of the smoothing parameter. The smoothing parameter controls the trade-off between the model fit and the model smoothness. Then, when $\lambda \rightarrow \infty$ the fitted curve tends to a polynomial of degree $d - 1$, if the degree of the B -spline is equal to or higher to the penalty order, i.e. if $q \geq d$. When $\lambda = 0$, the result is a the least squares estimate in (2.3). Therefore, the estimation of the degree of smoothness for the model consists in the estimation of the smoothing parameter λ . We discuss the selection of the optimal amount of λ in next section. Figure 2.3 shows the fitted curves for different values of λ .

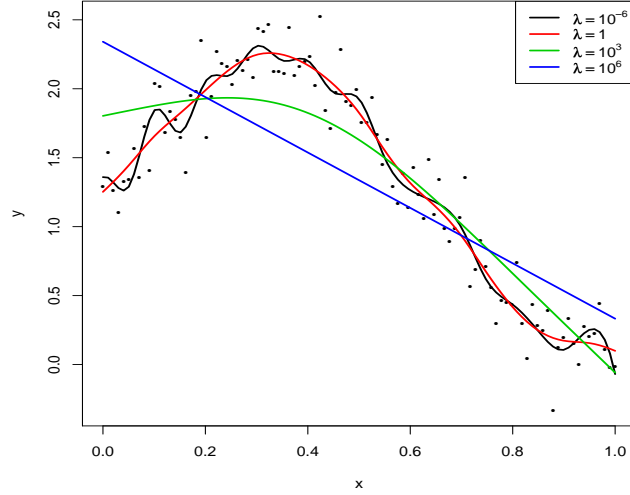


Figure 2.3: Fitted P -spline curves with different values of $\lambda = \{10^{-6}, 1, 10^3, 10^6\}$ and a second order penalty, $d = 2$.

2.1.2 Some basic definitions

P -splines have a number of statistical properties and results. In this Section, We some of the most useful.

Hat matrix

From [Equation 2.5](#), we can obtain the *smoother matrix*, or also called *hat-matrix* of the model for a given value of λ :

$$\mathbf{H} = \mathbf{B} (\mathbf{B}'\mathbf{B} + \mathbf{P})^{-1} \mathbf{B}' . \quad (2.7)$$

The matrix \mathbf{H} yields the fitted values $\hat{\mathbf{y}}$, such that $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, and therefore that for a given λ , the smoother is linear. It is also a extremely useful tool since it gives a measure of the effective dimension of the model.

Effective degrees of freedom

[Hastie and Tibshirani \(1990\)](#) defined the *effective dimension* of a smoother, or the *effective degrees of freedom* as the trace of the hat-matrix. Using the properties of the trace, we have that for comfortable matrices \mathbf{A} and \mathbf{B} , $\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$. A computationally

more efficient way of calculating the effective dimension is:

$$\text{ED} = \text{trace}(\mathbf{H}) = \text{trace}((\mathbf{B}'\mathbf{B} + \mathbf{P})^{-1}\mathbf{B}'\mathbf{B}), \quad (2.8)$$

which involves the calculations of $c \times c$ matrices, instead of the matrix \mathbf{H} of size $n \times n$. The value of the trace of \mathbf{H} is, hence, determined by the size of basis \mathbf{B} and the amount of smoothing. The value of ED lies between the number of columns of the B -spline basis, c , when $\lambda = 0$ and the order of the penalty d , when $\lambda \rightarrow \infty$. It is also important to note that, as a difference with respect to the classical linear regression, in P -splines, \mathbf{H} is not a projection matrix, since it is not idempotent, i.e. $\mathbf{H}^2 \neq \mathbf{H}$.

Confidence intervals and standard errors

The approximate variance of the fitted curve $\mathbf{B}\hat{\boldsymbol{\theta}}$ is given by

$$\text{Var}(\mathbf{B}\hat{\boldsymbol{\theta}}) \approx \sigma^2 \mathbf{B}(\mathbf{B}'\mathbf{B} + \mathbf{P})^{-1}\mathbf{B}'. \quad (2.9)$$

The diagonal elements of (2.9) are useful to construct twice standard error bands for the fitted curve. These intervals are the same as proposed in Wahba (1983) and Nychka (1988) from a bayesian perspective. (See Ruppert et al., 2003, Chapter 6, for more details).

The standard errors for the fitted curve can be calculated by analogy to linear regression as follows

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{B}\hat{\boldsymbol{\theta}}\|^2}{n - \text{ED}}. \quad (2.10)$$

2.1.3 Smoothing parameter selection

As addressed in Section 2.1.1, the P -spline model fit requires the choice of the amount of the penalty over the regression coefficients, in other words, a optimal choice of the smoothing parameter λ . There exists several methods to choose the optimal value of λ . We can classify the smoothing parameter selection methods in two main groups: methods based on cross-validation and methods based on an information criterion.

Cross-validation methods

The idea of cross-validation methods is to leave-out one observation in turn and then fit the model to the remaining data and calculate the squared difference between the missing data and its prediction, i.e. $\sum_{i=1}^n (y_i - \hat{y}_{-i})^2$, (see Stone, 1974). However, instead of repeating the process n times, a more efficient alternative is given by the fact that (see Hastie and Tibshirani, 1990, pg. 43):

$$y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{1 - h_{ii}}, i = 1, \dots, n$$

where h_{ii} are the diagonal elements of the hat matrix \mathbf{H} in (2.7). Therefore, we define the *ordinary cross-validation* as:

$$\text{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2, \quad (2.11)$$

A modified version of the CV criteria is the *generalized cross-validation* criteria which has some advantages over CV (as discussed in Craven and Wahba (1979); Wahba (1990)). For this criteria we compute:

$$\text{GCV}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - \sum_{i=1}^n h_{ii}/n} \right)^2, \quad (2.12)$$

where $\sum_{i=1}^n h_{ii}$, is the trace of the hat matrix defined in (2.8). Then the optimal λ minimizes the expressions in (2.11) and (2.12). Both CV and GCV criteria have some potential drawbacks, as for example: the computation becomes expensive when several smoothing parameters are considered. However, numerically stable and efficient methods have been proposed in the literature (See Wood, 2004, for discussion).

Information criteria methods

The other group of methods for smoothing parameter selection, are those based on information criterion. The idea is to compromise the goodness of fit and the complexity of a model by the correction of the log-likelihood of a fitted model for the effective dimension. Eilers and Marx (1996) suggest to minimize the information criterion (IC):

$$\text{IC} = \text{Dev}(\mathbf{y}; \boldsymbol{\theta}, \lambda) + \delta \text{ED}(\boldsymbol{\theta}, \lambda). \quad (2.13)$$

The deviance (Dev) is a measure of the quality of the fit, and it is defined as:

$$\text{Dev}(\mathbf{y}, \hat{\mathbf{y}}) = 2 \{ \mathcal{L}(\mathbf{y}) - \mathcal{L}(\hat{\mathbf{y}}) \}, \quad (2.14)$$

where $\mathcal{L}(\cdot)$ denotes the log-likelihood function. For Gaussian data, as it is the case at this point, the deviance is simply the residual sum of squares $\sum_{i=1}^n (y_i - \hat{y}_i)^2$.

For non-Gaussian data, the deviance is based on a generalizations of the sum of squares, and given the distributional assumptions it may take different expressions. The term δ penalizes the effective dimension of the model (ED) in (2.8). When $\delta = 2$

and $\delta = \log(n)$, we have the Akaike information criterion (AIC) (Akaike, 1973) and the Bayesian information criteria (BIC) (Schwarz, 1978), respectively. The BIC penalizes the model complexity more heavily than the AIC, specially when n is large. Improved versions of the AIC were discussed in Hurvich and Simonoff (1998).

2.1.4 P -splines for non-Gaussian responses

The P -spline methodology can be extended to the case of non-Gaussian data under the generalized linear models (GLMs) framework (Nelder and Wedderburn, 1972). GLMs expand the linear model for response distributions other than normal in an unified approach. The main reference is the book by McCullagh and Nelder (1989). Let \mathbf{y} the vector of responses and \mathbf{X} be a corresponding design matrix. The one-parameter exponential family model, with canonical link, is characterized by the joint density

$$f(\mathbf{y}; \boldsymbol{\beta}) = \exp \{ \mathbf{y}'(\mathbf{X}\boldsymbol{\beta}) - \mathbf{1}'b(\mathbf{X}\boldsymbol{\beta}) + \mathbf{1}'c(\mathbf{y}) \} \quad (2.15)$$

where $\boldsymbol{\beta}$ is the vector of coefficients. The log-likelihood of $\boldsymbol{\beta}$ is

$$\mathcal{L}(\boldsymbol{\beta}) = \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \mathbf{1}'b(\mathbf{X}\boldsymbol{\beta}) - \mathbf{1}'b(\mathbf{X}\boldsymbol{\beta}) + \mathbf{1}'c(\mathbf{y}) \quad (2.16)$$

The exponential family includes many distribution such as Normal, Poisson, Binomial or Gamma, they can be expressed in the form:

$$f(\mathbf{y}; \boldsymbol{\eta}) = \exp \left(\frac{\mathbf{y}\boldsymbol{\eta} - b(\boldsymbol{\eta})}{\phi} + c(\mathbf{y}, \phi) \right), \quad (2.17)$$

for some functions $b(\boldsymbol{\eta})$ and $c(\mathbf{y}, \phi)$, and where ϕ is the dispersion parameter. It can be shown that $\mathbb{E}[\mathbf{y}] = b'(\boldsymbol{\eta})$ and $\text{Var}[\mathbf{y}] = \phi \cdot b''(\boldsymbol{\eta})$, where $b(\boldsymbol{\eta})$ and $b''(\boldsymbol{\eta})$ are the first and second derivatives of b . The basic structure of a GLM is:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}), \text{ and } \boldsymbol{\mu} = \mathbb{E}(\mathbf{y}) = g^{-1}(\boldsymbol{\eta}), \quad (2.18)$$

where $\boldsymbol{\eta}$ is named as the *linear predictor*, and g is a monotonic differentiable function which relates the mean with the linear predictor, and therefore called *link function*. In addition, a GLM requires the choice of a distribution (within the exponential family). In the Gaussian case, we have that $\boldsymbol{\eta} = \boldsymbol{\mu}$, and thus, the link function is the identity. There are many choices of link functions and usually the *canonical link* is selected.

The P -spline methodology for a generalized linear model (P -GLM), can be easily extended (see Eilers and Marx, 1996). The linear predictor is defined as $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\theta}$, with basis \mathbf{B} and vector of coefficients $\boldsymbol{\theta}$, already defined in Section 2.1. The estimation in

GLMs uses the iterative reweighted least squares (IRLS) method, which can be easily adapted to a P -GLM. Now, the penalty is subtracted from the log-likelihood to form the *penalized log-likelihood* function:

$$\mathcal{L}_p(\boldsymbol{\theta}; \mathbf{y}) = \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) - \frac{1}{2} \boldsymbol{\theta}' \mathbf{P} \boldsymbol{\theta}, \quad (2.19)$$

where $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ is the ordinary log-likelihood function, and \mathbf{P} is the penalty matrix. Maximizing (2.19) we obtain the system of equations:

$$\mathbf{B}'(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{P} \boldsymbol{\theta},$$

which yields the penalized version of the scoring algorithm:

$$(\mathbf{B}' \tilde{\mathbf{W}} \mathbf{B} + \mathbf{P}) \hat{\boldsymbol{\theta}} = \mathbf{B}' \tilde{\mathbf{W}} \mathbf{B} \tilde{\boldsymbol{\theta}} + \mathbf{B}'(\mathbf{y} - \tilde{\boldsymbol{\mu}}), \quad (2.20)$$

where, $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\theta}}$ and $\tilde{\mathbf{W}}$ denotes the current approximate solution, and $\hat{\boldsymbol{\theta}}$ denotes the updated estimate of $\boldsymbol{\theta}$. The matrix \mathbf{W}_δ is diagonal with elements

$$w_{ii} = \frac{1}{v_i} \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2,$$

where v_i is the variance of y_i given μ_i . The algorithm (2.20), can be written as:

$$(\mathbf{B}' \tilde{\mathbf{W}} \mathbf{B} + \mathbf{P}) \hat{\boldsymbol{\theta}} = \mathbf{B}' \tilde{\mathbf{W}} \tilde{\mathbf{z}}, \quad (2.21)$$

where $\tilde{\mathbf{z}} = \tilde{\boldsymbol{\eta}} + \tilde{\mathbf{W}}^{-1}(\mathbf{y} - \tilde{\boldsymbol{\mu}})$ is known as the *working vector*. The estimating IRLS algorithm is summarized as follows:

-
- 1: set an initial value of $\boldsymbol{\theta}$ ($= \hat{\boldsymbol{\theta}}_{old}$)
 - 2: use $\hat{\boldsymbol{\theta}}_{old}$ to estimate \mathbf{W} and $\boldsymbol{\mu}_{old}$
 - 3: let $\hat{\boldsymbol{\eta}}_{old} = \mathbf{B} \hat{\boldsymbol{\theta}}_{old}$, get the \mathbf{z}_{new}
 - 4: obtain the new estimate $\hat{\boldsymbol{\theta}}_{new}$
 - 5: repeat 2 to 4 until convergence.
-

It follows from (2.21) that we can obtain the hat-matrix for a P -spline GLM as

$$\mathbf{H} = \mathbf{B}(\mathbf{B}' \hat{\mathbf{W}} \mathbf{B} + \mathbf{P})^{-1} \mathbf{B}' \hat{\mathbf{W}}, \quad (2.22)$$

where $\hat{\mathbf{W}}$ is the weight matrix at convergence. As addressed in (2.8), the effective dimension (ED) of a GLM is then the trace of the hat-matrix (2.22). The smoothing parameter in a P -GLM can be selected with methods described in Section 2.1.3, except that

for the information criteria the definition of the deviance in (2.14), depends on the generalization of the sum of squares in the GLM. An example of GLMs includes Poisson regression which models count data using the Poisson distribution. We use the Poisson case for illustrative purposes, and also because we will review the modelling of count data with P -splines for the spatial case in Chapter 3.

P -GLM for count data

We start by introducing an example that will serve to illustrative regression models for count data. The model commonly used is the Poisson, although other assumptions on the distribution can be considered as Geometric or Negative Binomial (for a extensive discussion on count data regression see Cameron and Trivedi (1998)). To illustrate the methodology for count data, we analyze the observed deaths of the female greek population in year 1960 (Kostaki and Panousis, 2001).

The Poisson P -GLM is constructed as follows. Let be \mathbf{y} the response vector of *deaths*, and \mathbf{x} the regressor variable *age-at-death*, where $\mathbf{x}' = (20, 21, \dots, 84)'$. We have the linear predictor with log-link given by:

$$\boldsymbol{\eta} = \log(\boldsymbol{\mu}) = \mathbf{B}\boldsymbol{\theta}, \quad (2.23)$$

where $\mathbf{B}\boldsymbol{\theta}$ represents the P -spline regression basis and coefficients, and $\boldsymbol{\mu} = \exp(\mathbf{B}\boldsymbol{\theta})$. In the Poisson case, the dispersion parameter $\phi = 1$, and the diagonal matrix of weights is $\mathbf{W} = \text{diag}(\boldsymbol{\mu})$ and the canonical link is the log.

Figure 2.4 shows the fitted P -spline curve with a B -spline basis constructed within 15 knots, a cubic spline and a second order penalty. The smoothing parameter was chosen by BIC, since for smoothing mortality data with P -splines, as suggested by Currie et al. (2004), the AIC tends to undersmooth. The BIC is defined as $\text{BIC} = \text{Dev} + \log n \text{ ED}$, where in the case of Poisson data, the *deviance* (Dev) is

$$\text{Dev} = \sum_{i=1}^n \left\{ y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right\}, \quad (2.24)$$

and ED is the effective dimension. The optimal value of the smoothing parameter chosen by BIC was $\lambda = 39.81$. This level of smoothing reduced the degrees of freedom from 18 (the number of fitted parameters, i.e. number of columns of \mathbf{B}) to an effective dimension of about 10.

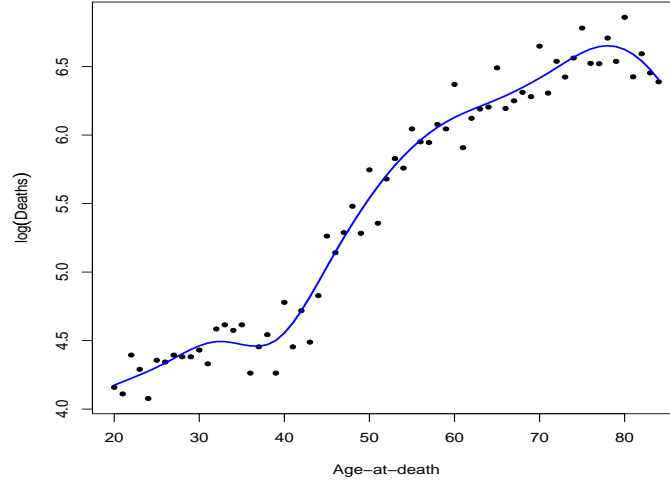


Figure 2.4: Fitted Poisson P -GLM to Greeks data, with optimized $\lambda = 39.81$ by BIC.

Bayesian P -splines

Brezger and Lang (2003) and Lang and Brezger (2004) extended the P -spline methodology to the Bayesian framework. From this perspective, the difference matrix D is replaced by its stochastic analog: first order differences corresponds to a first-order random walk and second order differences to a second-order random walk. Thus, we have for a B -spline basis B , of $n \times c$, we have c coefficients θ_j as first and second-order random walks, defined as:

$$\theta_j = \theta_{j-1} + v_j, \quad j = 2, \dots, c \quad y \quad (2.25)$$

$$\theta_j = 2\theta_{j-1} - \theta_{j-2} + v_j, \quad j = 3, \dots, c \quad (2.26)$$

where v_j are Gaussian errors $v_j \sim \mathcal{N}(0, \tau^2)$. Equations in (2.25) and (2.26) are then rewritten as:

$$\begin{aligned} \theta_j | \theta_{j-1} &\sim \mathcal{N}(\theta_{j-1}, \tau^2) \quad \text{and} \\ \theta_j | \theta_{j-1}, \theta_{j-2} &\sim \mathcal{N}(2\theta_{j-1} - \theta_{j-2}, \tau^2) \end{aligned}$$

The amount of smoothing is then controlled by the parameter τ^2 , which corresponds to $\tau^2 = \sigma^2/\lambda$ in the classical approach. Thus the priors in (2.25) and (2.26) are:

$$\theta_j | \tau^2 \propto \exp \left(-\frac{1}{2\tau^2} \theta' P \theta \right), \quad (2.27)$$

where the rank of the penalty matrix P is $c - q$, where q is the order of the random walk ($q = 1, 2$), thus the prior (2.27) is improper (see Brezger and Lang, 2003; Lang

and Brezger, 2004; Brezger and Lang, 2008, for further details). Bayesian P -splines are implemented in the software `BayesX` (Brezger et al., 2005).

Once we have presented the main aspects and definitions of the P -spline methodology, in next Section, we show that P -splines can be formulated as a mixed model. We based the mixed model representation in a reparameterization of the model basis, this lead us to the standard mixed model equations, and estimation methods. We will use this representation for the rest of the chapters.

2.2 Penalized splines and mixed models

Linear mixed effects models, or simply *mixed models*, are an extension of regression models which incorporate *random effects*. Several authors (Speed, 1991; Wang, 1998a; Zhang et al., 1998; Brumback and Rice, 1998; Verbyla et al., 1999) have addressed the connection between smoothing splines and mixed models. Wang (1998b) considered the mixed model formulation with correlated errors, and Lin and Zhang (1999) introduced the generalized additive mixed models for additive models and non-Gaussian responses.

The interest on this representation is due to the possibility of including smoothing in a large class of models (from correlated data to longitudinal studies and survival analysis), and the use of the methodology and software already developed for mixed models for estimation and inference. In the P -spline context, several authors have extended the model formulation into a mixed model (see Brumback et al., 1999; Coull et al., 2001b; Wand, 2002, among others). However, these authors used truncated polynomials as regression bases. We use the original B -spline basis of Eilers and Marx (1996) and follow a similar approach as in Currie and Durbán (2002); Currie et al. (2006) and Durbán et al. (2006).

Mixed models

The standard mixed model formulation is given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \text{ with } \boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{G}) \text{ and } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{\Lambda}), \quad (2.28)$$

where \mathbf{X} and \mathbf{Z} are the model matrices and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the *fixed* and *random* effects coefficients respectively. The random effects have covariance matrix \mathbf{G} , which depends on a variance of the random effects σ_α^2 . We will assume that the errors are *i.i.d.*, and then $\mathbf{\Lambda}$ is the identity matrix. We will focus on the general formulation in (2.28), for a more extensive review of mixed models as grouped data, longitudinal studies, multilevel

data or repeated measurements data, see [Searle et al. \(1992\)](#); [Verbeke and Molenberghs \(2000\)](#); [Pinheiro and Bates \(2000\)](#); [McCulloch and Searle \(2001\)](#) among others.

Estimation

From model (2.28), under the assumption of normality and *i.i.d.* errors, the marginal distribution of \mathbf{y} is normal with mean $\mathbf{X}\beta$ and variance:

$$\mathbf{V} = \sigma^2 \mathbf{I} + \mathbf{ZGZ}' \quad (2.29)$$

the log-likelihood of (2.28) is:

$$\mathcal{L}(\beta, \alpha, \sigma_\alpha^2, \sigma^2) = -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta),$$

where the variance components are included through the variance matrix (2.29). At fixed $(\sigma_\alpha^2, \sigma^2)$, if we take derivatives of the log-likelihood $\mathcal{L}(\cdot)$ with respect to β and α , and set them to zero, we find the estimates of the coefficients as:

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad (2.30)$$

$$\hat{\alpha} = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}), \quad (2.31)$$

(see [Pawitan, 2001](#), Chapter 17, for details).

However, it is known that the maximum likelihood estimates are biased, since they do not take into account the degrees of freedom used for the fixed effects estimation. An alternative estimation method which explicitly accounts for this loss of degrees of freedom is the *restricted maximum likelihood estimation* (see [Patterson and Thompson, 1971](#); [Harville, 1974](#); [Schall, 1991](#)). Then, deriving the profile the log-likelihood, it is possible to estimate the variance components by maximizing the residual maximum log-likelihood (REML), $\mathcal{L}_R(\sigma_\alpha^2, \sigma^2)$:

$$-\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} \log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2} \mathbf{y}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}. \quad (2.32)$$

For computational efficiency, it is possible to avoid the direct calculation the determinant and inverse of the variance component matrix \mathbf{V} , of dimension $n \times n$. Given that \mathbf{V} is an example of a Schur complement, it can be shown that:

$$|\mathbf{V}| = \sigma^{2n} |\mathbf{G}| |\mathbf{G}^{-1} + \frac{1}{\sigma^2} \mathbf{Z}'\mathbf{Z}| \quad (2.33)$$

$$\mathbf{V}^{-1} = \frac{1}{\sigma^2} (\mathbf{I} - \mathbf{Z}(\sigma^2 \mathbf{G}^{-1} + \mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}') \quad (2.34)$$

(see Searle et al., 1992). Note that, both expression (2.33) and (2.34), involves the inverses of $m \times m$ matrices, where m in this case denotes the number of random effects coefficients, and it is smaller than n .

2.2.1 Mixed models representation of P -splines

One of the many attractive features of the mixed model formulation of a spline model, is that the smoothing parameter, becomes the ratio between the variance of the residuals and the variance of the random effects, i.e. $\lambda = \sigma^2 / \sigma_\alpha^2$ (See Brumback et al., 1999; Ruppert et al., 2003). And therefore, the selection of the smoothing parameter becomes a variance components estimation problem instead of the optimization of a cross-validation method or an information criteria.

The aim is to reformulate the P -spline model into a mixed model (2.28). This reformulation can be viewed as a reparameterization of the original non-parametric model, for which we transform the model B -spline basis into a new model basis, i.e.:

$$B \rightarrow [X : Z].$$

This representation decomposes the fitted values as the sum of a polynomial/unpenalized part ($X\beta$) and a non-linear/penalized ($Z\alpha$) smooth term. There are several alternatives depending on the bases and the penalty used. We follow the approach by Currie and Durbán (2002) and Currie et al. (2006), and use the B -spline basis and the usual penalty P to reparameterize the original model into a mixed model.

Lemma 2.1. *The transformation matrix T to reparameterize the model basis and coefficients of a P -spline model into a mixed model representation is given by:*

$$T = [T_n : T_s], \quad (2.35)$$

where T is an orthogonal matrix, with submatrices T_n and T_s , which contain respectively the eigenvectors of the null and non-null part of the singular value decomposition of the penalty $D'D$.

Proof of orthogonality of T . Let $D'D = U\Sigma U'$, be the singular value decomposition of the penalty matrix, the matrix of eigenvalues U can be splitted in two parts:

$$U = [U_n : U_s], \quad (2.36)$$

where U_n contains the null part (of dimension $c \times q$) and U_s contains the span or the non-null part of the decomposition (of dimension $c \times (c - q)$). The diagonal matrix Σ

contains the eigenvalues of the SVD of $D'D$, with q null eigenvectors. Then, we can decompose the penalty as

$$D'D = [U_n : U_s] \begin{bmatrix} \mathbf{0}_q & \\ & \tilde{\Sigma} \end{bmatrix} \begin{bmatrix} U'_n \\ U'_s \end{bmatrix},$$

where $\mathbf{0}_q$ is square matrix of zeroes of order q and $\tilde{\Sigma}$ are the $(c - q)$ positive eigenvalues. For the uni-dimensional case, we take $T = [U_n : U_s]$, and then the matrix T is orthogonal since, we have:

$$TT' = U_n U'_n + U_s U'_s = UU' = I_c,$$

where q is the penalty order, and c is the number of columns of the B -spline basis B . ■

Lemma 2.2. *Given the transformation matrix T in (2.35), such that $BT = [X : Z]$. The fixed and random effects matrices are:*

$$X = BU_n, \text{ and} \tag{2.37}$$

$$Z = BU_s, \tag{2.38}$$

and the new coefficients by:

$$\beta = U'_n \theta \quad \text{and} \quad \alpha = U'_s \theta.$$

Proof. Immediate. ■

Theorem 2.1. *Given the orthogonal transformation matrix T in (2.35), and the penalty matrix $P = \lambda D'D$. The mixed model block-diagonal penalty is given by:*

$$\Phi = T'PT = \text{blockdiag}(\mathbf{0}_q, F), \text{ with } F = \lambda \tilde{\Sigma}, \tag{2.39}$$

where $\mathbf{0}_q$ is a square matrix of zeroes of order equal to the number of fixed effects q , and $\tilde{\Sigma}$ is the diagonal matrix with diagonal elements equal to the non-zero eigenvalues of the SVD of P .

Proof. Since $\begin{pmatrix} \beta \\ \alpha \end{pmatrix} = T'\theta$ and T is orthogonal, then:

$$T \begin{pmatrix} \beta \\ \alpha \end{pmatrix} = TT'\theta = \theta.$$

Then, the penalty $\theta' P \theta$ becomes:

$$\lambda \theta' D' D \theta \rightarrow \lambda (\beta' : \alpha') T' D' D T \begin{pmatrix} \beta \\ \alpha \end{pmatrix},$$

so then, the new penalty is $\lambda T D' D T$, but given the orthogonal matrix T in (2.35), we have

$$\begin{aligned} \lambda T D' D T &= \lambda \begin{bmatrix} U_n' \\ U_s' \end{bmatrix} D' D \begin{bmatrix} U_n \\ U_s \end{bmatrix} = \\ &= \lambda \begin{pmatrix} U_n' D' D U_n & U_n' D' D U_s \\ U_s' D' D U_n & U_s' D' D U_s \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{0}_q & \\ & \tilde{\Sigma} \end{pmatrix} = \begin{pmatrix} \mathbf{0}_q & \\ & \lambda \tilde{\Sigma} \end{pmatrix} \end{aligned}$$

where $U_n' D' D U_n = \mathbf{0}_q$ and $\tilde{\Sigma} = U_s' D' D U_s$, we obtain the result in (2.39). ■

Corollary 2.1. *Given (2.39), the fixed effects β are unpenalized, only the random effects, α , are penalized by the diagonal matrix F , that contains the positive eigenvalues of the SVD of $D' D$. Then, given the transformation matrix $T = [T_n : T_s]$, the mixed model penalty can be constructed as:*

$$F = T_s' P T_s, \quad (2.40)$$

where T_s is the submatrix of eigenvectors corresponding to the non-zero eigenvalues.

Proof. For the uni-dimensional case, $T_s = U_s$. Then:

$$F = \lambda U_s D' D U_s' = \lambda \tilde{\Sigma},$$

where $\tilde{\Sigma}$ is the diagonal matrix of non-zero eigenvalues of the SVD of $D' D$. ■

Corollary 2.2. *The original penalty matrix P , can be recovered from the new penalty F , by:*

$$P = T_s F T_s'. \quad (2.41)$$

Proof. We have seen in Theorem 2.1, we have that: $\Phi = T' P T'$, then:

$$T \Phi T' = T T' P T T' = P, \text{ and}$$

$$P = [T_n : T_s] \begin{pmatrix} \mathbf{0}_q & \\ & F \end{pmatrix} \begin{bmatrix} T_n' \\ T_s' \end{bmatrix} = T_s F T_s'. \quad \text{■}$$

Since the fixed parameters β 's are unpenalized, the fixed effect matrix $X = BU_n$, may be replaced by any sub-matrix such that:

- The composed matrix $[X : Z]$ has full rank. This also implies that both X and Z have full column rank.
- X and Z are orthogonal, i.e. $X'Z = 0$.

Assuming a second order penalty, i.e. $q = 2$, the diagonal matrix of eigenvalues has two zeroes and $c - 2$ positive eigenvalues. Then, the fixed effects matrix can be taken as:

$$X = [\mathbf{1} : x], \quad (2.42)$$

where $\mathbf{1}$ is a vector of ones and x is the covariate vector. Then, when $\lambda \rightarrow \infty$ the null model is the polynomial part $X\beta$ (of degree $q - 1$), and the random part can be considered as deviates from the null model. In [Section 2.3](#) we will see the extension of the methodology to the multidimensional case.

Given the new basis and the new penalty, the penalized sum of squares (2.4) becomes:

$$S(\beta, \alpha; \lambda) = (y - X\beta - Z\alpha)'(y - X\beta - Z\alpha) + \alpha'F\alpha. \quad (2.43)$$

Taking derivatives on (2.43) with respect to the parameters, it is straightforward to obtain the standard mixed model equations in (2.30) and (2.31). Now, with the reparameterization, the variance components matrix defined as $G = \sigma^2 F^{-1}$.

Hat matrix and confidence intervals in smoothing mixed models

Given the new transformed basis and penalty, the hat-matrix and its trace is calculated as:

$$H = [X : Z] \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1} \end{bmatrix}^{-1} [X : Z]', \quad (2.44)$$

for which

$$\text{trace}(H) = \text{trace} \left\{ \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix} \right\}. \quad (2.45)$$

Both expressions (2.45) and (2.44) can be computed taking advantages of the symmetry of the cross-products involved, as for instance: $X'X$, $X'Z$, and $Z'Z$. The mixed model software available (as `lme` function in R and `PROC MIXED` procedure in SAS[®]) includes efficient code for the fitting of mixed models.

The confidence intervals for the smooth curve $f(x)$, can be calculated using the estimated curve $\hat{f}(x) = \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\alpha}$. The variability of the estimated curve depends on whether the randomness of α is taken into account or not. It is possible to argue that the formulation of a P -spline as a mixed model, α is a device used to model curvature (see [Green, 1999](#)), and then ϵ , takes into account the variability around the curve.

[Ruppert et al. \(2003\)](#) suggest the use of a bias adjusted confidence intervals given using:

$$\text{Var}(\hat{f}(x) - f(x)) = \sigma^2 \mathbf{H}, \quad (2.46)$$

where are those obtained by [Hastie and Tibshirani \(1986\)](#), and where \mathbf{H} is the hat-matrix in (2.44). From a bayesian perspective [Wahba \(1983\)](#) and [Nychka \(1988\)](#), obtained bayesian confidence intervals for smoothing splines which are equivalent to the bias adjusted intervals using (2.46).

2.2.2 P -splines as generalized linear mixed models

The extension of the GLMs framework to include random effects is known as *generalized linear mixed models* (GLMMs). We consider the Poisson P -GLM, to show how random effects can be incorporated into a P -GLMM. As we showed in Section 2.1.4, in the P -GLM case, we have the linear predictor $\eta = \mathbf{B}\theta$. From the smoothing mixed model approach, we have that $\mathbf{B}\theta = \mathbf{X}\beta + \mathbf{Z}\alpha$, and therefore, the joint density (2.15) becomes:

$$f(\mathbf{y}|\alpha) = \exp \left\{ \mathbf{y}'(\mathbf{X}\beta + \mathbf{Z}\alpha) - \mathbf{1}' \exp(\mathbf{X}\beta + \mathbf{Z}\alpha) - \mathbf{1}' \log(\Gamma(\mathbf{y} + 1)) \right\}, \quad (2.47)$$

where $\alpha \sim \mathcal{N}(0, \mathbf{G})$ and $\eta = \mathbf{X}\beta + \mathbf{Z}\alpha$.

A full likelihood analysis in GLMMs usually involves numerical integration techniques to evaluate (2.47). [Breslow and Clayton \(1993\)](#) popularized the use of *Penalized Quasilikelihood* (PQL) methods developed by [Stiratelli et al. \(1984\)](#) and [Schall \(1991\)](#) for estimation and inference in these models. PQL is a very simple method for estimation of GLMMs, it can be easily implemented by iterative fitting a linear mixed model to a modified dependent variable. The PQL estimates are obtained of the coefficients (β, α) considering the random effects α as fixed parameters, and penalizing the likelihood according to the distribution of α . Thus, for given values of the variance components $(\sigma^2, \sigma_\alpha^2)$, and density (2.47), the parameters (β, α) are obtained by maximizing the *penalized log-likelihood*:

$$\log \{f(\mathbf{y}|\alpha)\} - \frac{1}{2} \alpha' \mathbf{G}^{-1} \alpha. \quad (2.48)$$

Maximization of (2.48) leads us to the score equations:

$$\mathbf{X}'(\mathbf{y} - \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})) = 0 \quad (2.49)$$

$$\mathbf{Z}'(\mathbf{y} - \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})) = \mathbf{G}^{-1}\boldsymbol{\alpha}. \quad (2.50)$$

The system of equations in (2.49) and (2.50) can be solved using a Fisher's scoring algorithm with *working vector* $\mathbf{z} = \boldsymbol{\eta} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu})$. Thus, the coefficients are given by:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{z} \quad (2.51)$$

$$\hat{\boldsymbol{\alpha}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad (2.52)$$

with $\mathbf{V} = \mathbf{W}^{-1} + \mathbf{Z}\mathbf{G}\mathbf{Z}'$, and diagonal matrix of weights $\mathbf{W} = \text{diag}(\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}))$ in the Poisson case. Then, conditional on the estimates obtained in (2.51) and (2.52), the variance components are estimated by the approximate REML quasi-likelihood, $QL(\boldsymbol{\beta}, \sigma^2, \sigma_\alpha^2)$:

$$-\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} \log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2} \mathbf{z}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{z} \quad (2.53)$$

The PQL solution is obtained by iteration between (2.51), (2.52) and (2.53) until convergence. It is possible to avoid $n \times n$ matrix evaluations in the iterative procedure, using:

$$\mathbf{V}^{-1} = \mathbf{W} - \mathbf{W}\mathbf{Z}(\mathbf{G}^{-1} + \mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W} \quad (2.54)$$

$$|\mathbf{V}| = |\mathbf{W}|^{-1} |\mathbf{G}| |\mathbf{G}^{-1} + \mathbf{Z}'\mathbf{W}\mathbf{Z}|. \quad (2.55)$$

The PQL solution is only an approximation to a full likelihood analysis, except in the Gaussian GLMM, where it is exact. Sometimes the approximation works remarkably well (as in the Poisson case), but in some situations (e.g. logistic regression) the variance components may not be estimated correctly. We will use the PQL approach in Chapter 3 Section 3.4.1 for the estimation in Poisson spatial count data in the presence of overdispersion.

2.3 Multidimensional smoothing with P-splines

In the previous sections, we explained in detail the smoothing mixed model approach for a regression model with a single covariate. Now, we consider a general non-

parametric k -dimensional regression model:

$$\mathbf{y} = f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2), \quad (2.56)$$

where f is a smooth function of the k -regressors, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_k)'$, of lengths n_1, n_2, \dots , and n_k , respectively. Most of the smoothing methods for multivariate data are constrained to the *curse of the dimensionality* (Bellman, 1961), in the sense that the cost of computing and store requirements depends exponentially on the dimensionality k . This is one of the reasons why the use of smoothing splines in several dimensions becomes computationally prohibitive.

The natural extension, in the context of the splines literature, is the use of *Tensor products* of the regression bases (de Boor, 1978). P -splines are low-rank smoothers, so they present a computational advantage in for the multidimensional case with respect to smoothing splines. We extend the smoothing mixed model methodology following the approach of Currie et al. (2006) and Eilers et al. (2006) for the multidimensional case. They proposed efficient algorithms for smoothing with P -splines when data present an array structure and we will adapt them to be used in the mixed model context (see also Wood, 2006b,a, for a similar approach).

In the multidimensional smoothing context, the data can fall into two categories: (i) they can come as large grids of values (as for example: mortality life-tables usually are classified by age-at-death and year-of-death or image data) or (ii) they can be irregular or scattered data (as for example spatial data). The extension of the P -spline methodology to the multidimensional case, requires the construction of a regression B -spline basis that will depend on the type of data structure considered. In this Section we will focus on the first case of smoothing data on multidimensional grids, and will discuss the case of scattered data in Chapter 3.

2.3.1 Smoothing multidimensional data with array structure

Suppose we are interested in fitting model (2.56), and assume that:

$$f(\mathbf{x}_1, \dots, \mathbf{x}_k) = \mathbf{B}\boldsymbol{\theta}, \quad (2.57)$$

where \mathbf{B} is the B -spline regression basis, and $\boldsymbol{\theta}$ the vector of coefficients. When data have an array structure, we define \mathbf{Y} , of dimension $n_1 \times n_2 \times \dots \times n_k$, as the response k -dimensional array, and $\mathbf{y} = \text{vec}(\mathbf{Y})$, the vector of length $n_1 n_2 \dots n_k \times 1$. The smooth multidimensional surface is constructed from the Tensor product of the individual or

marginal B -spline basis for each covariate. Then, the basis for model (2.56) is

$$\mathbf{B} = \mathbf{B}_k \otimes \cdots \otimes \mathbf{B}_2 \otimes \mathbf{B}_1,$$

where symbol \otimes is the Kronecker product of two matrices, and $\mathbf{B}_i = \mathbf{B}(x_i)$ is the marginal B -spline basis for each x_i , and $i = 1, \dots, k$.

For simplicity, we will illustrate the bivariate case ($k = 2$), i.e. $f(x_1, x_2)$, where we have the two regressors:

$$\begin{aligned} \mathbf{x}_1 &= (x_{1i}, \dots, x_{1n_1})' \quad \text{and} \\ \mathbf{x}_2 &= (x_{2j}, \dots, x_{2n_2})' \quad \text{for } i = 1, \dots, n_1 \text{ and } j = 1, \dots, n_2. \end{aligned}$$

The vector \mathbf{y} of length $n \times 1$, where $n = n_1 n_2$, can be arrange into a matrix of n_1 rows and n_2 columns, as:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n_2} \\ y_{21} & y_{22} & \cdots & y_{2n_2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n_1 1} & \cdots & \cdots & y_{n_1 n_2} \end{bmatrix}_{n_1 \times n_2}.$$

The regression basis for smoothing is

$$\mathbf{B} = \mathbf{B}_2 \otimes \mathbf{B}_1, \tag{2.58}$$

where $\mathbf{B}_1 = \mathbf{B}(x_1)$ and $\mathbf{B}_2 = \mathbf{B}(x_2)$, of dimensions $n_1 \times c_1$ and $n_2 \times c_2$ respectively. Then, the dimension of (2.58) is $n_1 n_2 \times c_1 c_2$.

Figure 2.5 shows the Kronecker product of two cubic B -spline basis. A full B -spline basis looks like Figure 2.6, with knots equally spaced over x_1 and x_2 domains. For illustrative purposes, we only considered a small portion of basis functions. The regression coefficients are placed on the peaks the “hills” and smoothness of the fitted surface is ensured by imposing a penalty over the coefficients $\boldsymbol{\theta}$ in both directions.

A bivariate P -spline model can be written as:

$$f(x_1, x_2) = \mathbf{B}\boldsymbol{\theta} = (\mathbf{B}_2 \otimes \mathbf{B}_1)\boldsymbol{\theta}. \tag{2.59}$$

The $c_1 c_2 \times 1$ elements of the vector of coefficients $\boldsymbol{\theta}$ can be arranged into a matrix $\boldsymbol{\Theta}$ of

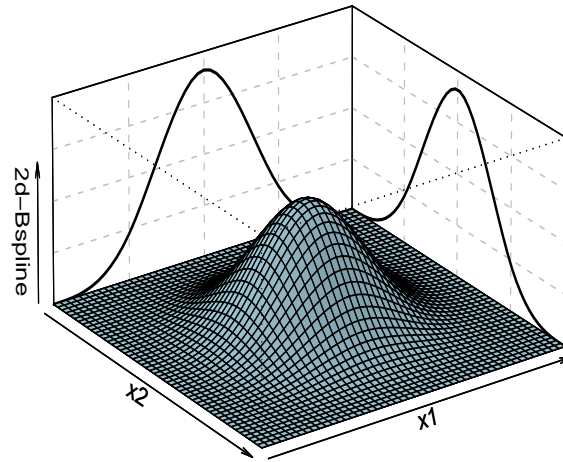


Figure 2.5: Tensor product of two cubic splines.

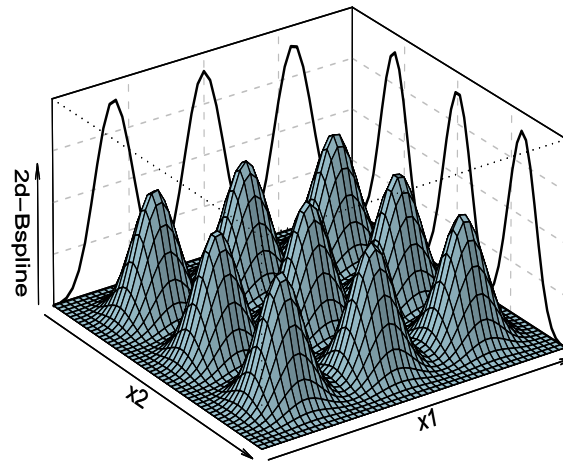


Figure 2.6: A portion of the full basis, consisting in the Tensor product of nine cubic splines.

dimensions $c_1 \times c_2$, and $\text{vec}(\Theta) = \theta$:

$$\Theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1c_2} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2c_2} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{c_1 1} & \cdots & \cdots & \theta_{c_1 c_2} \end{bmatrix}.$$

Currie et al. (2006) and Eilers et al. (2006) developed an arithmetic of arrays which allows to smooth data over multidimensional grids. These algorithms are extended to the GLM framework, so they refer to them as *generalized linear array models* or GLAMs.

The essence of GLAM is to arrange the vectors into matrices/arrays, i.e. the expression in (2.59), can be written as:

$$f(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{B}\boldsymbol{\theta} \equiv \mathbf{B}_1\boldsymbol{\Theta}\mathbf{B}_2'. \quad (2.60)$$

It follows that expression (2.60) is equivalent to:

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} B_i(\mathbf{x}_1) B_j(\mathbf{x}_2) \boldsymbol{\theta}_{ij},$$

where $B_i(\mathbf{x}_1)$ and $B_j(\mathbf{x}_2)$ are the i^{th} and j^{th} columns of the marginal B -spline basis. Note that, the result in (2.60) is a $n_1 \times n_2$ matrix, that can be easily vectorized into the vector of length $n_1 n_2 \times 1$, with the $\text{vec}(\cdot)$ operator.

Following the ideas of the unidimensional case, we penalized the coefficients vector $\boldsymbol{\theta}$ by a penalty matrix \mathbf{P} . The penalty in two dimensions penalizes rows and columns of the matrix $\boldsymbol{\Theta}$ of coefficients. The appropriate penalty on rows of $\boldsymbol{\Theta}$ is:

$$\sum_{i=1}^{c_1} \boldsymbol{\theta}_i' \mathbf{D}_1' \mathbf{D}_1 \boldsymbol{\theta}_i = \boldsymbol{\theta}' (\mathbf{D}_1' \mathbf{D}_1 \otimes \mathbf{I}_{c_1}) \boldsymbol{\theta}, \quad (2.61)$$

and similarly on the columns, i.e.:

$$\sum_{j=1}^{c_1} \boldsymbol{\theta}_j' \mathbf{D}_2' \mathbf{D}_2 \boldsymbol{\theta}_j = \boldsymbol{\theta}' (\mathbf{I}_{c_1} \otimes \mathbf{D}_2' \mathbf{D}_2) \boldsymbol{\theta}, \quad (2.62)$$

where \mathbf{D}_1 and \mathbf{D}_2 are the differences matrices acting on the rows and columns of $\boldsymbol{\Theta}$, defined in (2.3.1).

Finally, we obtain the penalty matrix \mathbf{P} in two dimensions as:

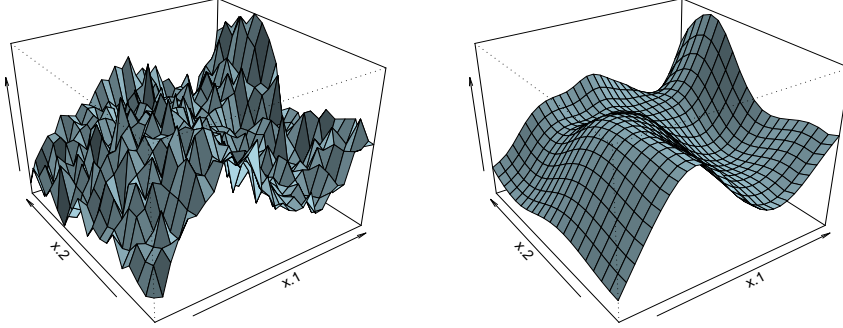
$$\mathbf{P} = \lambda_1 \underbrace{\mathbf{I}_{c_2} \otimes \mathbf{D}_1' \mathbf{D}_1}_{\mathbf{P}_1} + \lambda_2 \underbrace{\mathbf{D}_2' \mathbf{D}_2 \otimes \mathbf{I}_{c_2}}_{\mathbf{P}_2}, \quad (2.63)$$

where λ_1 and λ_2 are the smoothing parameters for each dimension of the model. Note that, this penalty allow for an *anisotropic* smoothing, since it considers a different amount of smoothing in each dimension ($\lambda_1 \neq \lambda_2$). The expression $\mathbf{P}_1 + \mathbf{P}_2$ is known as a *Kronecker sum*, and can also be written as:

$$\mathbf{P} = \lambda_2 \mathbf{D}_2' \mathbf{D}_2 \oplus \lambda_1 \mathbf{D}_1' \mathbf{D}_1.$$

Figure 2.7 illustrates the P -spline smoothing for array data in two dimensions. The raw

(a) Raw and smoothed perspective plots



(b) Raw and smoothed image plots

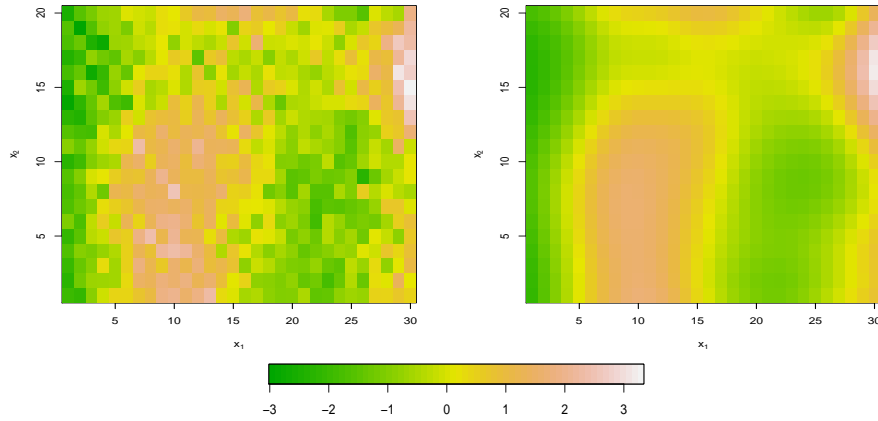


Figure 2.7: Raw data and smoothed data with 8-by-5 knots and

data consists in $n = 600$ observations with covariates x_1, x_2 of lengths $n_1 = 30$ and $n_2 = 20$. The estimation of these multidimensional models is done, using the same methodology we showed in previous sections, using cross validation or AIC/BIC for the estimation of the smoothing parameters λ_1 and λ_2 . The usual strategy is to evaluate the criteria for a grid values of the smoothing parameters.

However, smoothing in several dimensions is susceptible to runaway problems with storage and computational time. The GLAM algorithms developed in [Currie et al. \(2006\)](#) and [Eilers et al. \(2006\)](#), reduce the computational time and can be implemented in standard software as R or MATLAB[®]. In [Section A.2 of Appendix A](#), we present some of the useful array arithmetic multiplications and their implementation in R code. We will see an application of the penalty (2.77) for a three-dimensional model in [Chapter 4](#).

2.3.2 Multidimensional mixed models representation of P-splines

In this Section, we extend the results presented in Section 2.2 to the multidimensional case. We show how it is possible to extend the reparameterization into a mixed model in several dimensions, and use the GLAM arithmetic for a fast and efficient implementation. Our aim is to reformulate the multidimensional model in (2.57) as a mixed model:

$$f(x_1, \dots, x_d) = \mathbf{X}\beta + \mathbf{Z}\alpha, \quad \text{with } \alpha \sim \mathcal{N}(0, \mathbf{G}),$$

where the basis and coefficients are reparameterized as:

$$\mathbf{B} \rightarrow [\mathbf{X} : \mathbf{Z}] \quad \text{and} \quad \boldsymbol{\theta} \rightarrow (\beta, \alpha).$$

We consider the two-dimensional case (as presented in Section 2.3.1) with regression basis $\mathbf{B} = \mathbf{B}_2 \otimes \mathbf{B}_1$, and penalty matrix \mathbf{P} defined in (2.63). Recall that, the mixed model reparameterization proposed in Section 2.2, consists in applying the singular value decomposition on the penalty matrix \mathbf{P} . In two dimensions, the SVD over the Kronecker sum $\mathbf{P}_1 + \mathbf{P}_2$, allows the simultaneous diagonalization of the penalty matrix \mathbf{P} (see Horn and Johnson, 1991, for details), as a function of the individual penalty diagonalizations: $\mathbf{D}'_1 \mathbf{D}_1 = \mathbf{U}_1 \boldsymbol{\Sigma}_1 \mathbf{U}'_1$ and $\mathbf{D}'_2 \mathbf{D}_2 = \mathbf{U}_2 \boldsymbol{\Sigma}_2 \mathbf{U}'_2$, with matrices \mathbf{U}_1 and \mathbf{U}_2 , defined as in (2.36), and $\tilde{\boldsymbol{\Sigma}}_1$ and $\tilde{\boldsymbol{\Sigma}}_2$ has $(c_1 - q_1)$ and $(c_2 - q_2)$ positive eigenvalues respectively.

Following a similar procedure as in the univariate case, we need to find a transformation matrix such that we reparameterize model bases and coefficients into a mixed model.

Lemma 2.3. *The transformation matrix \mathbf{T} to reparameterize the model bases and coefficients in a two-dimensional case into a mixed model is a partitioned matrix defined by:*

$$\mathbf{T} = [\underbrace{\mathbf{U}_{2n} \otimes \mathbf{U}_{1n}}_{\mathbf{T}_n} : \underbrace{\mathbf{U}_{2s} \otimes \mathbf{U}_{1n} : \mathbf{U}_{2n} \otimes \mathbf{U}_{1s} : \mathbf{U}_{2s} \otimes \mathbf{U}_{1s}}_{\mathbf{T}_s}], \quad (2.64)$$

where \mathbf{T}_n is the block corresponding to the fixed part and \mathbf{T}_s the blocks for the random part. The matrix \mathbf{T} is an orthogonal matrix of dimension $n_1 n_2 \times c_1 c_2$.

Proof. Let us define, the marginal transformation matrices $\mathbf{T}_1 = [\mathbf{U}_{1n} : \mathbf{U}_{1s}]$ and $\mathbf{T}_2 = [\mathbf{U}_{2n} : \mathbf{U}_{2s}]$, of dimensions $n_2 \times c_2$ and $n_1 \times c_1$, respectively. The Kronecker product of both partitioned matrices, $\mathbf{T}_2 \otimes \mathbf{T}_1$, is the $n_1 n_2 \times c_1 c_2$ matrix obtained by replacing each

block of T_2 with the Kronecker product of U_{2n} and U_{2s} by T_1 , that is:

$$\begin{aligned} T_2 \otimes T_1 &= [U_{2n} : U_{2s}] \otimes T_1 = [U_{2n} \otimes T_1 : U_{2s} \otimes T_1] = \\ &= (U_{2n} \otimes [U_{1n} : U_{1s}] : U_{2s} \otimes [U_{1n} : U_{1s}]). \end{aligned} \quad (2.65)$$

We reorder the block matrices in (2.65), into fixed (T_n) and random (T_s) parts, as shown in (2.64). Then,

$$\begin{aligned} TT' &= U_{2n}U_{2n}' \otimes U_{1n}U_{1n}' + U_{2s}U_{2s}' \otimes U_{1n}U_{1n}' + U_{2n}U_{2n}' \otimes U_{1s}U_{1s}' + U_{2s}U_{2s}' \otimes U_{1s}U_{1s}' = \\ &= (U_{2n}U_{2n}' + U_{2s}U_{2s}') \otimes U_{1n}U_{1n}' + (U_{2n}U_{2n}' + U_{2s}U_{2s}') \otimes U_{1s}U_{1s}' = \\ &= (U_{2n}U_{2n}' + U_{2s}U_{2s}') \otimes (U_{1n}U_{1n}' + U_{1s}U_{1s}') = I_{c_1} \otimes I_{c_2} = I_{c_1 c_2}. \end{aligned}$$

And thus T is orthogonal. ■

Lemma 2.4. *Given the transformation matrix T in (2.64) such that, $BT = [X : Z]$. The fixed and random effects matrices are:*

$$X = X_2 \otimes X_1, \text{ and} \quad (2.66)$$

$$Z = (Z_2 \otimes X_1 : X_2 \otimes Z_1 : Z_2 \otimes Z_1), \quad (2.67)$$

where $X_k = B_k U_{kn}$ and $Z_k = B_k U_{ks}$, for $k = 1, 2$.

Proof. By Property A.6 of the mixed product rule of Kronecker products (see Appendix A.1). We take the fixed effects matrix as:

$$\begin{aligned} X &= BT_n = (B_2 \otimes B_1)(U_{2n} \otimes U_{1n}) = \\ &= (B_2 U_{2n} \otimes B_1 U_{1n}) = X_2 \otimes X_1. \end{aligned} \quad (2.68)$$

And random effects matrix as:

$$\begin{aligned} Z &= BT_s = (B_2 \otimes B_1)(U_{2n} \otimes U_{1s} : U_{2s} \otimes U_{1n} : U_{2s} \otimes U_{1s}) = \\ &= (B_2 U_{2s} \otimes B_1 U_{1n} : B_2 U_{2n} \otimes B_1 U_{1s} : B_2 U_{2s} \otimes B_1 U_{1s}) = \\ &= (Z_2 \otimes X_1 : X_2 \otimes Z_1 : Z_2 \otimes Z_1). \end{aligned} \quad (2.69)$$

The new coefficients are:

$$\beta = T_n' \theta, \quad \text{and} \quad \alpha = T_s' \theta.$$

■

Remark 2.1. Note that, both X and Z have a Kronecker product structure, so it is possible to use the array algorithms (we will see the implementation of GLAM algorithms in mixed models in next Section).

Theorem 2.2 (Mixed model penalty F in two dimensions). *Given the orthogonal transformation matrix T in two dimensions defined in (2.64) and the penalty matrix in (2.63). By 2.1, the mixed model block-diagonal penalty is $F = T'_s P T_s$, that for the two-dimensional case, is given by:*

$$F = \begin{pmatrix} \lambda_2 \tilde{\Sigma}_2 \otimes I_{q_1} & & \\ & \lambda_1 I_{q_2} \otimes \tilde{\Sigma}_1 & \\ & & \lambda_1 I_{c_2-q_2} \otimes \tilde{\Sigma}_1 + \lambda_2 \tilde{\Sigma}_2 \otimes I_{c_1-q_1} \end{pmatrix}, \quad (2.70)$$

where $\tilde{\Sigma}_1$, and $\tilde{\Sigma}_2$, of dimensions $(c_1 - q_1) \times (c_1 - q_1)$ and $(c_2 - q_2) \times (c_2 - q_2)$, are the diagonal matrices of positive eigenvalues of $D'_1 D_1$ and $D'_2 D_2$.

Proof. Given the definition of matrix T in (2.64), and the penalty matrix P , expressed as a Kronecker sum $P_1 + P_2$, we can obtain the mixed model penalty for the two-dimensional case in (2.39) as:

$$\begin{aligned} \Phi &= T' P T = T' (\lambda_1 P_1 + \lambda_2 P_2) T = \lambda_1 T' P_1 T + \lambda_2 T' P_2 T = \\ &= \lambda_1 \begin{pmatrix} U'_{2n} \otimes U'_{1n} \\ U'_{2s} \otimes U'_{1n} \\ U'_{2n} \otimes U'_{1s} \\ U'_{2s} \otimes U'_{1s} \end{pmatrix} (I_{c_2} \otimes D'_1 D_1) (U_{2n} \otimes U_{1n} : U_{2s} \otimes U_{1n} : U_{2n} \otimes U_{1s} : U_{2s} \otimes U_{1s}) + \\ &+ \lambda_2 \begin{pmatrix} U'_{2n} \otimes U'_{1n} \\ U'_{2s} \otimes U'_{1n} \\ U'_{2n} \otimes U'_{1s} \\ U'_{2s} \otimes U'_{1s} \end{pmatrix} (D'_2 D_2 \otimes I_{c_1}) (U_{2n} \otimes U_{1n} : U_{2s} \otimes U_{1n} : U_{2n} \otimes U_{1s} : U_{2s} \otimes U_{1s}) = \\ &= \lambda_1 \begin{pmatrix} U'_{2n} U_{2n} \otimes U'_{1n} D'_1 D_1 U_{1n} & & & \\ & U'_{2s} U_{2s} \otimes U'_{1n} D'_1 D_1 U_{1n} & & \\ & & U'_{2n} U_{2n} \otimes U'_{1s} D'_1 D_1 U_{1s} & \\ & & & U'_{2s} U_{2s} \otimes U'_{1s} D'_1 D_1 U_{1s} \end{pmatrix} + \\ &+ \lambda_2 \begin{pmatrix} U'_{2n} D'_2 D_2 U_{2n} \otimes U'_{1n} U_{1n} & & & \\ & U'_{2s} D'_2 D_2 U_{2s} \otimes U'_{1n} U_{1n} & & \\ & & U'_{2n} D'_2 D_2 U_{2n} \otimes U'_{1s} U_{1s} & \\ & & & U'_{2s} D'_2 D_2 U_{2s} \otimes U'_{1s} U_{1s} \end{pmatrix} = \\ &= \lambda_1 \begin{pmatrix} I_{q_2} \otimes \mathbf{0}_{q_1} & & & \\ & I_{c_2-q_2} \otimes \mathbf{0}_{q_1} & & \\ & & I_{c_2-q_2} \otimes \tilde{\Sigma}_1 & \\ & & & I_{c_2-q_2} \otimes \tilde{\Sigma}_1 \end{pmatrix} + \lambda_2 \begin{pmatrix} \mathbf{0}_{q_2} \otimes I_{q_1} & & & \\ & \tilde{\Sigma}_2 \otimes I_{q_1} & & \\ & & \mathbf{0}_{q_2} \otimes I_{c_1-q_1} & \\ & & & \tilde{\Sigma}_2 \otimes I_{c_1-q_1} \end{pmatrix}. \quad (2.71) \end{aligned}$$

Given that, $U'_{2n}U_{2n} = I_{q_2}$ and $U'_{2s}U_{2s} = I_{c_2-q_2}$, and also that

$$U'_{2n}D'_2D_2U_{2n} = \mathbf{0}_{q_2} \quad \text{and} \quad U'_{1n}D'_1D_1U_{1n} = \mathbf{0}_{q_1},$$

we have that (2.71) becomes:

$$\Phi = \text{blockdiag}(\mathbf{0}_{q_1q_2}, \mathbf{F}),$$

where $\mathbf{0}_{q_1q_2}$ is a square matrix of zeroes of order q_1q_2 . Then, \mathbf{F} is the block-diagonal penalty matrix over the random effects coefficients α . ■

Given the definitions of the mixed model matrices \mathbf{X} and \mathbf{Z} , and the new mixed model penalty \mathbf{F} , we have obtained all the basic elements of a mixed model. The estimation of the coefficients and the variance components are done as shown in Section 2.2. In next Section, we will detail the array computations in the mixed model formulation using the GLAM algorithms.

An important result of the reparameterization shown above, is that the transformed penalty and model matrices, lead us to a very interesting decomposition of the fitted values. Note that, the diagonal matrix \mathbf{F} in (2.70) has three blocks, the first block involves the smoothing parameter λ_1 and the non-zero eigenvalues of D'_1D_1 , the second block has the smoothing parameter λ_2 and the non-zero eigenvalues of D'_2D_2 , and finally, the last block involves both smoothing parameters and a Kronecker sum of the non-zero eigenvalues.

In the case of a second order penalty in both dimensions, we can take $\mathbf{X}_1 = [\mathbf{1}_1 : \mathbf{x}_1]$ and $\mathbf{X}_2 = [\mathbf{1}_2 : \mathbf{x}_2]$, where $\mathbf{1}_1$ and $\mathbf{1}_2$ are column vectors of ones of length n_1 and n_2 respectively. Thus, the fixed effects matrices is given by:

$$\begin{aligned} \mathbf{X} &= [\mathbf{1}_1 : \mathbf{x}_1] \otimes [\mathbf{1}_2 : \mathbf{x}_2] = \\ &= [\mathbf{1} : \mathbf{1}_2 \otimes \mathbf{x}_1 : \mathbf{x}_2 \otimes \mathbf{1}_1 : \mathbf{x}_2 \otimes \mathbf{x}_1], \end{aligned} \quad (2.72)$$

where $\mathbf{1}$ is a column vector of ones of length n_1n_2 , and the random effects matrix is:

$$\mathbf{Z} = \underbrace{(\mathbf{Z}_2 \otimes [\mathbf{1}_1 : \mathbf{x}_1])}_{(**)} : \underbrace{([\mathbf{1}_2 : \mathbf{x}_2] \otimes \mathbf{Z}_1 : \mathbf{Z}_2 \otimes \mathbf{Z}_1)}_{(*)}, \quad (2.73)$$

where $(*)$ is equal to $[\mathbf{1}_2 \otimes \mathbf{Z}_1 : \mathbf{x}_2 \otimes \mathbf{Z}_1]$. In the case of $(**)$, the Kronecker product gives the same elements as $[\mathbf{Z}_2 \otimes \mathbf{1}_1 : \mathbf{Z}_2 \otimes \mathbf{x}_1]$, but the columns are in a different order.

Thus, the matrix Z can be written as

$$Z \equiv [Z_2 \otimes \mathbf{1}_1 : \mathbf{1}_2 \otimes Z_1 : x_2 \otimes Z_1 : Z_2 \otimes x_1 : Z_2 \otimes Z_1], \quad (2.74)$$

where symbol \equiv denotes that matrices (2.73) and (2.74) have same elements but in a different order. This partition of the model matrices allow us to represent the fitted surface as a sum of three terms (up to a constant term) i.e.:

(i) A marginal term for x_1 :

$$f_1(x_1) \equiv [\mathbf{1}_2 \otimes x_1 : \mathbf{1}_2 \otimes Z_1].$$

(ii) A marginal term for x_2 :

$$f_2(x_2) \equiv [x_2 \otimes \mathbf{1}_1 : Z_2 \otimes \mathbf{1}_1].$$

(iii) An interaction term for both x_1 and x_2 :

$$f_{1,2}(x_1, x_2) \equiv [x_2 \otimes x_1 : Z_2 \otimes x_1 : x_2 \otimes Z_1 : Z_2 \otimes Z_1].$$

This partition leads us to consider a decomposition of the two-dimensional surface as:

$$f(x_1, x_2) = f_1(x_1) + f_2(x_2) + f_{1,2}(x_1, x_2), \quad (2.75)$$

where functions f_1 and f_2 can be interpreted as an ANOVA-type model as *main effects* of the covariates x_1 and x_2 , and $f_{1,2}$ is a two-dimensional interaction surface or *interaction effect* between x_1 and x_2 . This decomposition is strongly related to the work proposed by Gu (2002) and the *Smoothing Spline Analysis of Variance* (or SS-ANOVA) models. Figure 2.8 shows the decomposition of the two dimensional surface as an ANOVA-type model. Note that, this decomposition leaves unchanged the interpretation of the penalties, and allow us to represent the fitted surface as a sum of marginal terms and the interaction. We extend this decomposition in the P -spline context in Chapter 4.

Three-dimensional smoothing mixed models

The extension of the smoothing mixed model methodology to more than two dimensions is straightforward. For $k = 3$, the model is given by:

$$\mathbb{E}[\mathbf{y}] = f(x_1, x_2, x_3), \quad (2.76)$$

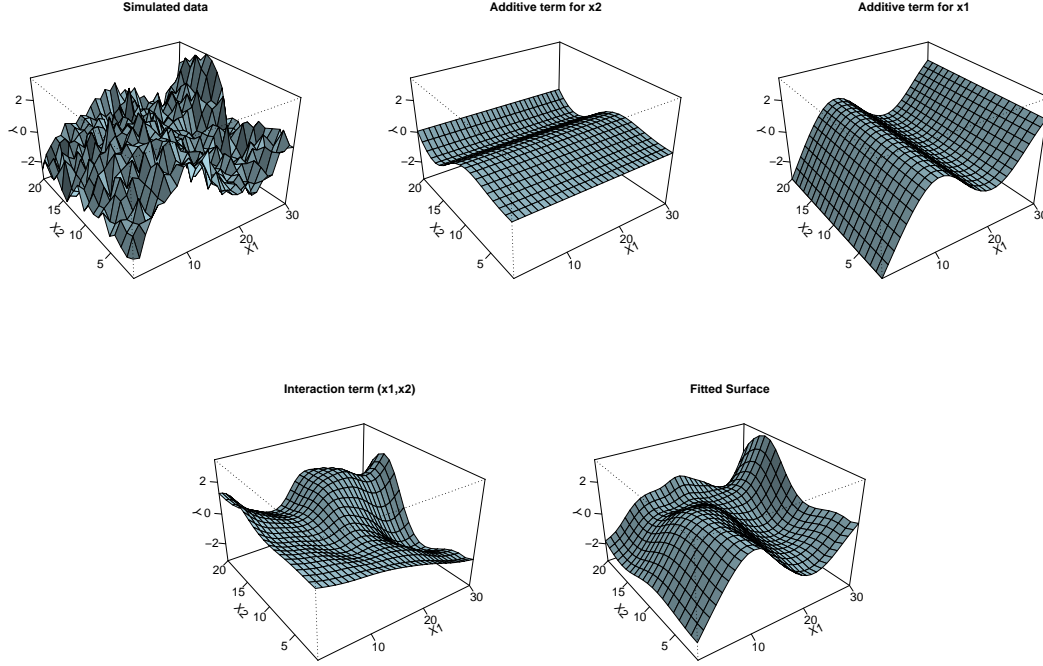


Figure 2.8: Decomposition of the two dimensional surface into additive terms and interactions.

with data Y arranged in a three-dimensional array of dimensions, $n_1 \times n_2 \times n_3$. The regression basis is $B = B_1 \otimes B_2 \otimes B_3$ of dimension $n_1 n_2 n_3 \times c_1 c_2 c_3$, and coefficients arranged in a three dimensional array Θ of dimension $c_1 \times c_2 \times c_3$ (as shown in [Figure 2.9](#)). The penalty matrix for the three-dimensional case is:

$$P = \lambda_1 D_1' D_1 \otimes I_{c_2} \otimes I_{c_3} + \lambda_2 I_{c_1} \otimes D_2' D_2 \otimes I_{c_3} + \lambda_3 I_{c_1} \otimes I_{c_2} \otimes D_3' D_3, \quad (2.77)$$

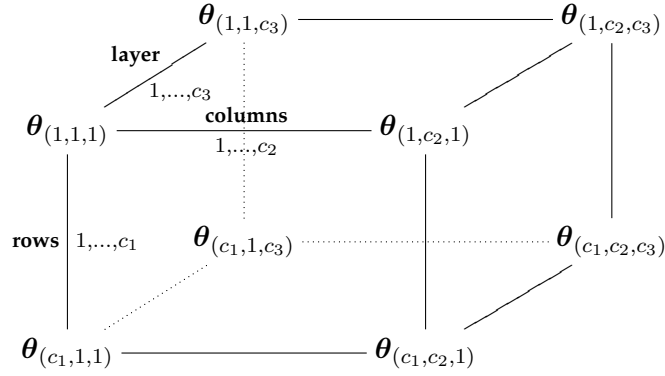
that can be written as a Kronecker sum as:

$$P = \lambda_1 D_1' D_1 \oplus \lambda_2 D_2' D_2 \oplus \lambda_3 D_3' D_3.$$

Now, we follow the same procedure as in [Section 2.3.2](#) for the bivariate case, and define the transformation matrix that reparameterizes the P -spline three-dimensional model into a mixed model.

The transformation matrix T can be defined as, the Kronecker product:

$$\begin{aligned} T &= \bigotimes_{k=1}^3 T_k = \bigotimes_{k=1}^3 [U_{kn} : U_{ks}] = \\ &= [U_{1n} : U_{1s}] \otimes [U_{2n} : U_{2s}] \otimes [U_{3n} : U_{3s}]. \end{aligned} \quad (2.78)$$

Figure 2.9: Array Θ of coefficients in three dimensions of $c_1 \times c_2 \times c_3$.

As in the two-dimensional case, we reorder the Kronecker product in (2.78), into two sub-blocks as $T = [T_n : T_s]$, where the first block T_n , corresponds to the fixed part:

$$T_n = [U_{1n} \otimes U_{2n} \otimes U_{3n}],$$

and the second block, T_s , to the random part:

$$T_s = [U_{1s} \otimes U_{2n} \otimes U_{3n} : U_{1n} \otimes U_{2s} \otimes U_{3n} : U_{1n} \otimes U_{2n} \otimes U_{3s} : \\ U_{1s} \otimes U_{2s} \otimes U_{3n} : U_{1s} \otimes U_{2n} \otimes U_{3s} : U_{1n} \otimes U_{2s} \otimes U_{3s} : \\ U_{1s} \otimes U_{2s} \otimes U_{3s}].$$

We reparameterize the regression basis into $BT = [X : Z]$, and obtain the fixed effects matrix

$$X = X_1 \otimes X_2 \otimes X_3, \quad (2.79)$$

and random effects matrix

$$Z = [Z_1 \otimes X_2 \otimes X_3 : X_1 \otimes Z_2 \otimes X_3 : X_1 \otimes X_2 \otimes Z_3 : \\ Z_1 \otimes Z_2 \otimes X_3 : Z_1 \otimes X_2 \otimes Z_3 : X_1 \otimes Z_2 \otimes Z_3 : \\ Z_1 \otimes Z_2 \otimes Z_3], \quad (2.80)$$

where $X_k = [\mathbf{1}_{n_k} : x_k]$ and $Z_k = B_k U_{ks}$, for $k = 1, 2, 3$.

The block-diagonal penalty matrix F for the three-dimensional model, is again given

by $\mathbf{F} = \mathbf{T}_s' \mathbf{P} \mathbf{T}_s$:

$$\mathbf{F} = \text{blockdiag}(\mathbf{F}_{(1)}, \mathbf{F}_{(2)}, \mathbf{F}_{(3)}, \mathbf{F}_{(1,2)}, \mathbf{F}_{(1,3)}, \mathbf{F}_{(2,3)}, \mathbf{F}_{(1,2,3)}), \quad (2.81)$$

where

$$\begin{aligned} \mathbf{F}_{(1)} &= \lambda_1 \tilde{\Sigma}_1 \otimes \mathbf{I}_{q_2} \otimes \mathbf{I}_{q_3}, \\ \mathbf{F}_{(2)} &= \lambda_2 \mathbf{I}_{q_1} \otimes \tilde{\Sigma}_2 \otimes \mathbf{I}_{q_3}, \\ \mathbf{F}_{(3)} &= \lambda_3 \mathbf{I}_{q_1} \otimes \mathbf{I}_{q_2} \otimes \tilde{\Sigma}_3, \\ \mathbf{F}_{(1,2)} &= \lambda_1 \tilde{\Sigma}_1 \otimes \mathbf{I}_{c_2-q_2} \otimes \mathbf{I}_{q_3} + \lambda_2 \mathbf{I}_{c_1-q_2} \otimes \tilde{\Sigma}_2 \otimes \mathbf{I}_{q_3}, \\ \mathbf{F}_{(1,3)} &= \lambda_1 \tilde{\Sigma}_1 \otimes \mathbf{I}_{q_2} \otimes \mathbf{I}_{c_3-2} + \lambda_3 \mathbf{I}_{c_1-q_1} \otimes \mathbf{I}_{q_2} \otimes \tilde{\Sigma}_3, \\ \mathbf{F}_{(2,3)} &= \lambda_2 \mathbf{I}_{q_1} \otimes \tilde{\Sigma}_2 \otimes \mathbf{I}_{c_3-q_3} + \lambda_3 \mathbf{I}_{q_1} \otimes \mathbf{I}_{q_2-c_2} \otimes \tilde{\Sigma}_3, \\ \mathbf{F}_{(1,2,3)} &= \lambda_1 \tilde{\Sigma}_1 \otimes \mathbf{I}_{c_2-q_2} \otimes \mathbf{I}_{c_3-q_3} + \lambda_2 \mathbf{I}_{c_1-q_1} \otimes \tilde{\Sigma}_2 \otimes \mathbf{I}_{c_3-q_3} + \lambda_3 \mathbf{I}_{c_1-q_1} \otimes \mathbf{I}_{c_1-2} \otimes \tilde{\Sigma}_3. \end{aligned}$$

Notice that, as we showed in the bivariate case, we can also decompose the three-dimensional smoothing, into marginal terms and interactions. The block structure of the model matrices \mathbf{X} and \mathbf{Z} in (2.79) and (2.80) and the block-diagonal terms of the penalty (2.81), lead us to identify each of the parts of the decomposition. We will discuss this model in Chapter 4.

2.3.3 Array methods for multidimensional smoothing mixed models

In this Section we show the use of GLAM algorithms in the smoothing mixed model context, for smoothing in multidimensional grids. In Section 2.3.2, we proposed the use of restricted or residual maximum likelihood (REML), for the estimation of the variance parameters. Given (2.32), and definitions of \mathbf{V} , $|\mathbf{V}|$ and \mathbf{V}^{-1} in (2.29), (2.33), and (2.34), we may use the GLAM algorithms for a fast and efficient computation of the matrix cross-products: $\mathbf{Z}'\mathbf{Z}$, $\mathbf{X}'\mathbf{Z}$, $\mathbf{X}'\mathbf{y}$ or $\mathbf{Z}'\mathbf{y}$, etc ... And estimate the variance components by REML.

To illustrate the implementation of the GLAM algorithm as mixed models, we divide the REML in several parts as:

$$-\frac{1}{2} \underbrace{\log |\mathbf{V}|}_{\text{part I}} - \frac{1}{2} \underbrace{\log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|}_{\text{part II}} - \frac{1}{2} \underbrace{(\mathbf{y}'\mathbf{V}^{-1}\mathbf{y})}_{\text{part III}} - \underbrace{\mathbf{y}'(\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y}}_{\text{part IV}}.$$

Using the GLAM notation in Currie et al. (2006) (see Appendix A.2.1 for basic definitions), for the mixed model formulation, we propose two additional definitions:

Definition 2.1 (\mathcal{A}_1 -form). Given the inner product $\mathbf{X}'\mathbf{W}_\delta\mathbf{X}$ and the k -dimensional array $\mathbf{X} = \mathbf{X}_k \otimes \dots \otimes \mathbf{X}_1$ of dimensions $n_i \times c_i$, for $i = 1, \dots, k$, where \mathbf{W}_δ is the matrix of

weights in a GLM, The \mathcal{A}_1 -form is the array:

$$\rho(\mathcal{G}(\mathbf{X}_k)', \dots, \rho(\mathcal{G}(\mathbf{X}_1)', \mathbf{W})), \text{ of dimension } c_1^2 \times \dots \times c_k^2, \quad (2.82)$$

where in the Gaussian case \mathbf{W} is a $n_1 \times n_2$ matrix of ones, i.e. $\mathbf{W} = \mathbf{1}\mathbf{1}'$. The resulting array must be reorganized into a square matrix $\mathbf{X}'\mathbf{W}_\delta\mathbf{X}$ of order $c_1c_2\dots c_k$.

Definition 2.2 (\mathcal{A}_2 -form). Given the product $\mathbf{X}'\mathbf{W}\mathbf{Z}$ and the k -dimensional array $\mathbf{X} = \mathbf{X}_k \otimes \dots \otimes \mathbf{X}_1$ of dimensions $n_i \times c_i$, and $\mathbf{Z} = \mathbf{Z}_k \otimes \dots \otimes \mathbf{Z}_1$ of dimensions $n_i \times p_i$, for $i = 1, \dots, k$. The \mathcal{A}_2 -form is

$$\rho(\mathcal{G}(\mathbf{Z}_k, \mathbf{X}_k)', \dots, \rho(\mathcal{G}(\mathbf{Z}_1, \mathbf{X}_1)', \mathbf{W})), \text{ of dimensions } c_1p_1 \times \dots \times c_kp_k. \quad (2.83)$$

The resulting array must be reorganized into the matrix $\mathbf{X}'\mathbf{W}_\delta\mathbf{Z}$ of dimension $c_1c_2\dots c_k \times p_1p_2\dots p_k$.

Now, we will detail the array computation of each part of the REML function. In order to simplified the notation, we consider the bivariate smoothing case for Gaussian data, and second order penalties (i.e. $q_1 = q_2 = 2$). Hence, the fixed effects matrix is given by:

$$\mathbf{X} = \mathbf{X}_2 \otimes \mathbf{X}_1, \text{ of dimension } n_1n_2 \times 4,$$

and random effects matrix:

$$\mathbf{Z} = (\mathbf{Z}_2 \otimes \mathbf{X}_1 : \check{\mathbf{Z}}_2 \otimes \mathbf{Z}_1), \text{ of dimension } n_1n_2 \times (c_1c_2 - 4),$$

where $\check{\mathbf{Z}}_2$ is the $n_1n_2 \times c_2$ partitioned matrix $\check{\mathbf{Z}}_2 = [\mathbf{X}_2 : \mathbf{Z}_2]$.

Part I: Array computation of $\log |\mathbf{V}|$

Given the variance components matrix $\mathbf{G} = \sigma^2\mathbf{F}^{-1}$, with \mathbf{F} a block-diagonal matrix defined in (2.70), the determinant of the variance of \mathbf{y} in (2.33), can be written as:

$$|\mathbf{V}| = \sigma^{2n} |\mathbf{F}^{-1}| |\mathbf{F} + \mathbf{Z}'\mathbf{Z}|. \quad (2.84)$$

In the two-dimensional case in (2.70), \mathbf{F} is a block-diagonal matrix with three blocks, i.e.:

$$\mathbf{F} = \text{blockdiag}(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3).$$

Then, \mathbf{F}^{-1} is the inverse of the diagonal elements of \mathbf{F} , and $|\mathbf{F}^{-1}|$ is the product of the inverse elements of the diagonal of \mathbf{F} .

For the computation of $\mathbf{Z}'\mathbf{Z}$, we have the block-symmetric matrix, formed by four blocks:

$$\mathbf{Z}'\mathbf{Z} = \begin{bmatrix} \mathbf{Z}'_2\mathbf{Z}_2 \otimes \mathbf{X}'_1\mathbf{X}_1 & \mathbf{Z}'_2\tilde{\mathbf{Z}}_2 \otimes \mathbf{X}'_1\mathbf{Z}_1 \\ \tilde{\mathbf{Z}}'_2\mathbf{Z}_2 \otimes \mathbf{Z}'_1\mathbf{X}_1 & \tilde{\mathbf{Z}}'_2\tilde{\mathbf{Z}}_2 \otimes \mathbf{Z}'_1\mathbf{Z}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{Z}'\mathbf{Z}_{(1)} & \mathbf{Z}'\mathbf{Z}_{(2)} \\ \mathbf{Z}'\mathbf{Z}_{(3)} & \mathbf{Z}'\mathbf{Z}_{(4)} \end{bmatrix},$$

of dimension $(c_1c_2 - 4) \times (c_1c_2 - 4)$. Blocks of $\mathbf{Z}'\mathbf{Z}$ can be computed using the \mathcal{A}_1 and \mathcal{A}_2 -forms, as:

- $\mathbf{Z}'\mathbf{Z}_{(1)}$ and $\mathbf{Z}'\mathbf{Z}_{(4)}$ are in the \mathcal{A}_1 -form:

$$\begin{aligned} \mathbf{Z}'_2\mathbf{Z}_2 \otimes \mathbf{X}'_1\mathbf{X}_1 &\equiv \rho(\mathcal{G}(\mathbf{Z}_2)', \rho(\mathcal{G}(\mathbf{X}_1)', \mathbf{W})), \\ \tilde{\mathbf{Z}}'_2\tilde{\mathbf{Z}}_2 \otimes \mathbf{Z}'_1\mathbf{Z}_1 &\equiv \rho(\mathcal{G}(\tilde{\mathbf{Z}}_2)', \rho(\mathcal{G}(\mathbf{Z}_1)', \mathbf{W})). \end{aligned}$$

- $\mathbf{Z}'\mathbf{Z}_{(2)}$ and $\mathbf{Z}'\mathbf{Z}_{(3)}$ are in the \mathcal{A}_2 -form:

$$\begin{aligned} \mathbf{Z}'_2\tilde{\mathbf{Z}}_2 \otimes \mathbf{X}'_1\mathbf{Z}_1 &\equiv \rho(\mathcal{G}(\tilde{\mathbf{Z}}_2, \mathbf{Z}_2)', \rho(\mathcal{G}(\mathbf{Z}_1, \mathbf{X}_1)', \mathbf{W})), \\ \tilde{\mathbf{Z}}'_2\mathbf{Z}_2 \otimes \mathbf{Z}'_1\mathbf{X}_1 &\equiv \rho(\mathcal{G}(\mathbf{Z}_2, \tilde{\mathbf{Z}}_2)', \rho(\mathcal{G}(\mathbf{X}_1, \mathbf{Z}_1)', \mathbf{W})). \end{aligned}$$

Note that, $\mathbf{Z}'\mathbf{Z}_{(2)}$ is the transpose of $\mathbf{Z}'\mathbf{Z}_{(3)}$, and viceversa, so we only need to calculate it once and transpose it.

To compute $|\mathbf{F} + \mathbf{Z}'\mathbf{Z}|$, we define the determinant by blocks as:

$$\left| \begin{pmatrix} \mathbf{F}_1 & \\ & \mathbf{F}^* \end{pmatrix} + \begin{pmatrix} \mathbf{Z}'\mathbf{Z}_{(1)} & \mathbf{Z}'\mathbf{Z}_{(2)} \\ \mathbf{Z}'\mathbf{Z}_{(3)} & \mathbf{Z}'\mathbf{Z}_{(4)} \end{pmatrix} \right| = \left| \begin{pmatrix} \mathbf{F}_1 + \mathbf{Z}'\mathbf{Z}_{(1)} & \mathbf{Z}'\mathbf{Z}_{(2)} \\ \mathbf{Z}'\mathbf{Z}_{(3)} & \mathbf{F}^* + \mathbf{Z}'\mathbf{Z}_{(4)} \end{pmatrix} \right| = \left| \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{pmatrix} \right|,$$

where $\mathbf{F}^* = \text{blockdiag}(\mathbf{F}_2, \mathbf{F}_3)$. The determinant of a block-diagonal partitioned matrix is:

$$\left| \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{C} \end{pmatrix} \right| = |\mathbf{A}| |\mathbf{C} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}|.$$

(See details in [Harville, 2000](#), pg. 189).

Finally:

$$\log |\mathbf{V}| = 2n \log(\sigma) + \log |\mathbf{F}^{-1}| + \log |\mathbf{F} + \mathbf{Z}'\mathbf{Z}|, \quad (2.85)$$

with $\log |\mathbf{F}^{-1}| = \sum \log(\mathbf{F}_{ii}^{-1})$, where \mathbf{F}_{ii}^{-1} are the inverse of the diagonal elements of the block-matrix \mathbf{F} .

Part II: Array computation of $|X'V^{-1}X|$

We rewrite the inverse of V as:

$$V^{-1} = \frac{1}{\sigma^2}(I - Z(F + Z'Z)^{-1}Z'), \quad (2.86)$$

Then, $X'V^{-1}X$ is given by:

$$X'V^{-1}X = \frac{1}{\sigma^2}(X'X - X'Z(F + Z'Z)^{-1}Z'X). \quad (2.87)$$

We need to compute: $X'X$, $Z'X$ and $X'Z$, i.e.:

$$X'X = (X_2 \otimes X_1)'(X_2 \otimes X_1) = X_2'X_2 \otimes X_1'X_1, \quad (2.88)$$

$$Z'X = \begin{bmatrix} (Z_2 \otimes X_1)' \\ (\tilde{Z}_2 \otimes Z_1)' \end{bmatrix} (X_2 \otimes X_1) = \begin{bmatrix} Z_2'X_2 \otimes X_1'X_1 \\ \tilde{Z}_2'X_2 \otimes Z_1'X_1 \end{bmatrix}. \quad (2.89)$$

Using the array arithmetic, we compute (2.88) and (2.89) as

$$X'X \equiv \rho(\mathcal{G}(X_2, Z_2)', \rho(\mathcal{G}(X_1, X_1)', W)), \quad \text{and} \quad Z'X \equiv \begin{bmatrix} \rho(\mathcal{G}(X_2, Z_2)', \rho(\mathcal{G}(X_1, X_1)', W)) \\ \rho(\mathcal{G}(X_2, \tilde{Z}_2)', \rho(\mathcal{G}(X_1, Z_1)', W)) \end{bmatrix},$$

where in the Gaussian case W is a $n_1 \times n_2$ matrix of ones, i.e. $W = \mathbf{1}\mathbf{1}'$.

Part III: Array computation of $y'V^{-1}y$

Given (2.86), we can write $y'V^{-1}y$ as:

$$y'V^{-1}y = \frac{1}{\sigma^2}(y'y - y'Z(F + Z'Z)^{-1}Z'y), \quad (2.90)$$

where $y'y$ is easily computed as sum of squares of the elements of y . Using the array methods $Z'y$ is computed as:

$$\begin{aligned} Z'y &= (Z_2 \otimes X_1 : \tilde{Z}_2 \otimes Z_1)'y = \begin{bmatrix} Z_2' \otimes X_1' \\ \tilde{Z}_2' \otimes Z_1' \end{bmatrix} y = \begin{bmatrix} (Z_2' \otimes X_1')y \\ (\tilde{Z}_2' \otimes Z_1')y \end{bmatrix} = \\ &\equiv \text{vec} \begin{bmatrix} X_1'Y Z_2 \\ Z_1'Y \tilde{Z}_2' \end{bmatrix} \equiv \text{vec} \begin{bmatrix} \rho(Z_2', \rho(X_1', Y)) \\ \rho(\tilde{Z}_2', \rho(Z_1', Y)) \end{bmatrix}. \end{aligned}$$

where Y is the $n_1 \times n_2$ response matrix, and $y'Z$ as the transpose of $Z'y$.

Part IV: Array computation of $\mathbf{y}' (\mathbf{V}^{-1} \mathbf{X} (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1}) \mathbf{y}$

We have already shown the computation of $\mathbf{X}' \mathbf{V}^{-1} \mathbf{X}$ in (2.87), in this part of the REML, we need its inverse and to compute: $\mathbf{X}' \mathbf{V}^{-1} \mathbf{y}$ and $\mathbf{y} \mathbf{V}^{-1} \mathbf{X}$. Given (2.86), $\mathbf{X}' \mathbf{V}^{-1} \mathbf{y}$ can be written as:

$$\mathbf{X}' \mathbf{V}^{-1} \mathbf{y} = \frac{1}{\sigma^2} (\mathbf{X}' \mathbf{y} - \mathbf{X}' \mathbf{Z} (\mathbf{F} + \mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y}), \quad (2.91)$$

where all the quantities were computed in previous sections, except to $\mathbf{X}' \mathbf{y}$, which is computed as:

$$\mathbf{X}' \mathbf{y} = (\mathbf{X}'_2 \otimes \mathbf{X}'_1) \mathbf{y} \equiv \rho(\mathbf{X}'_2, \rho(\mathbf{X}'_1, \mathbf{Y})).$$

The quantity $\mathbf{y} \mathbf{V}^{-1} \mathbf{X}$ is the transpose of $\mathbf{X}' \mathbf{V}^{-1} \mathbf{y}$.

Given the computations of each part, the REML in (2.32) can be fastly maximized to obtain the optimal values of the smoothing parameters $(\lambda_1, \dots, \lambda_k)$ and the variance σ^2 . The R functions and code to compute these formulas are shown in [Appendix A](#).

*"A picture is a poem without words".
Horace*

Chapter 3

Smoothing spatial data with penalized splines

In this Chapter, we adapt the methodology developed in [Chapter 2](#) and present the smoothing mixed approach as an unified framework for the analysis of spatial data. We have already given an extensive review of the methodology used in the univariate case, and the multidimensional setting in the case of data on array structure (or regular grids). However, spatial data structure is not regular and so, the Tensor product defined by the Kronecker product and the GLAM methodology cannot be directly applied. We propose the use of a different Tensor product of B -spline basis for this situation. The algebra required to extend the methodology to the mixed model framework is not straightforward, but we demonstrate that it is also possible to apply it using some matrix algebra results. We show some examples of spatial data smoothing in [Section 3.3](#). In particular, we apply the smoothing mixed model approach to the case of spatial count data, and extend this model to incorporate spatial random effects with a particular structure: we propose a hybrid model, called the "*Smooth-CAR model*", which combines a smooth model (modelled with a P -spline) for large scale variability (to capture the smooth spatial trend), and a conditionally autoregressive model (CAR) for the small scale variability. A simulation study is presented in [Section 3.5](#).

3.1 B -spline basis for spatial data

As we showed in [Chapter 2](#), P -spline regression depends on a regression basis and a penalty matrix which controls the smoothness of the fit. In the case of two-dimensional smoothing (as it is the case of spatial data) the election of basis and penalty is even more important, and the differences between the approaches are significant. Some authors

Table 3.1: Tensor products of two marginal B -splines basis for array and scattered data.

	Data structure	
	Regular grid/array	Scattered/spatial
Tensor Product	Kronecker	row-wise Kron.
Regression Basis	$B_2 \otimes B_1$	$B_2 \square B_1$
Basis dimension	$n_1 n_2 \times c_1 c_2$	$n \times c_1 c_2$

suggest the use of radial basis functions or Thin plate splines (Duchon, 1976), or a more computationally efficient version of Thin plate regression splines proposed by Wood (2003). These bases have the limitation of being isotropic smoothers, and the selection of knots to construct the basis is not trivial. We follow the B -spline basis approach with equally spaced knots presented in the previous Chapter.

The extension to the two-dimensional case depends on the structure of the data. If we consider scattered data, the basis is constructed from the Tensor product of marginal B -spline basis defined in Eilers et al. (2006) as the Box-Product or “row-wise” Kronecker product of the individual basis, denoted by symbol \square , and defined as:

$$B = B_2 \square B_1 = (B_2 \otimes \mathbf{1}'_{c_1}) \odot (\mathbf{1}'_{c_2} \otimes B_1), \quad (3.1)$$

where B_1 and B_2 are the B -spline basis along the longitude (x_1) and latitude (x_2) coordinates of dimensions $n \times c_1$ and $n \times c_2$. The basis B is of dimension $n \times c_1 c_2$, and the operator \odot is the *Hadamard* or “element-wise” matrix product and $\mathbf{1}_{c_1}$ and $\mathbf{1}_{c_2}$ are column vectors of ones of length c_1 and c_2 . The operations in (3.1) are such that row i of $B_2 \square B_1$ is the Kronecker product of the corresponding rows of B_2 and B_1 . The similar column-wise product is known as the *Khatri-Rao product* (Rao and Rao, 1998) (definitions and properties are given in Appendix B).

Table 3.1 summarizes the two types of products depending on the data structure. The dimension of the regression basis B is much more smaller (less number of rows) in the scattered case than in the regular grid case. Also, for the scattered or spatial case, both marginal bases have the same number of rows, since we have the same number, n , data points in x_1 and x_2 . It is important to note that the use of the row-wise Kronecker product implies that the GLAM structure is no longer available.

The regression basis in (3.1) allows us to smooth on spatial data, there is no restriction that spatial latitude and longitude coordinates should be on a regular grid. The extension of the P -spline methodology is then straightforward. The penalty over the regression coefficients θ is the same as in the two dimensional case shown in (2.63), since

the data structure does not affect the definition of the penalty matrix P on the rows and columns of the array of coefficients Θ . Fahrmeir and Lang (2001) used a penalty based on the Tensor product of the marginal penalties, defined by:

$$D_2' D_2 \otimes D_1' D_1. \quad (3.2)$$

However this type of penalty incurs into a rank deficiency problem (See discussion in Wood (2006b)). The rank of (3.2) is the product of the ranks of the marginal penalties, i.e. $(c_2 - q_2)(c_1 - q_1)$. A further drawback of using that penalty in spatial smoothing is the isotropy, given that (3.2) imposes the same amount of smoothing in both directions.

3.2 Mixed model reparameterization for spatial data

For spatial data, the P -spline model can be also reformulated as a multidimensional mixed model. The only difference with the case of data in an array structure shown in Section 2.3.2, is in the regression basis B . Now, the regression B -spline basis B is constructed by the Box-product defined in (3.1), instead of the Kronecker product of the marginal basis. Given that, the type of Tensor product depends on the data structure, and not on the coefficients (we can always arrange the coefficients in array form). We consider the same definition of the transformation matrix T for the two-dimensional case in (2.64), i.e.:

$$T = (\underbrace{U_{2n} \otimes U_{1n}}_{T_n} : \underbrace{U_{2s} \otimes U_{1n} : U_{2n} \otimes U_{1s} : U_{2s} \otimes U_{1s}}_{T_s}),$$

where the model basis is reparameterized as a mixed model, such that $BT = [X : Z]$. However, in the spatial case, the reparameterization into the model matrices X and Z , is not straightforward as in the data with array structure. As we showed in Section 2.3.2, we can obtain the fixed and random effects matrices as: $X = BT_n$ and $Z = BT_s$. In order to demonstrate the reparameterization, we use some matrix algebra results in Rao and Rao (1998) and Liu (1999, 2002), defined in Appendix B.

Theorem 3.1 (Mixed model bases for spatial data). *Let T be the transformation matrix for smoothing two-dimensional data defined in (2.64). The fixed and random effects matrices for the mixed model representation of P -splines are:*

$$X = X_2 \square X_1, \quad (3.3)$$

$$Z = (Z_2 \square X_1 : X_2 \square Z_1 : Z_2 \square Z_1). \quad (3.4)$$

Proof. Given spatial B -spline basis B in (3.1) and transformation matrix T , defined in (2.64). The fixed and random effects matrices are obtained as $X = BT_n$ and $Z = BT_s$. In the spatial case, we have:

$$X = BT_n = B(U_{2n} \otimes U_{1n}) = (B_2 \square B_1)(U_{2n} \otimes U_{1n}) \quad (3.5)$$

Let us denote by symbol $*$, as the *Khatri-Rao product* of two matrices (see definition in Appendix B), we use the result in Proposition B.3, and define

$$(B'_2 * B'_1) = (B_2 \square B_1)'. \quad (3.6)$$

By Proposition B.2, and using (3.6), we have:

$$(U'_{2n} \otimes U'_{1n})(B_2 \square B_1)' = (B_2 U_{2n} \square B_1 U_{1n})'. \quad (3.7)$$

Taking the transpose in both sides of (3.7), we have:

$$(B_2 \square B_1)(U_{2n} \otimes U_{1n}) = (B_2 \square B_1)(U'_{2n} \otimes U'_{1n})' = (B_2 U_{2n} \square B_1 U_{1n}). \quad (3.8)$$

Then, by result in (3.8), we obtain that the fixed effects matrix X in (3.5) is

$$X = B_2 U_{2n} \square B_1 U_{1n} = X_2 \square X_1,$$

where $X_k = B_k U_{kn}$, for $k = 1, 2$. For the random effects matrix, we have:

$$Z = BT_s = (B_2 \square B_1)(U_{2s} \otimes U_{1n} : U_{2n} \otimes U_{1s} : U_{2s} \otimes U_{1s}). \quad (3.9)$$

Using the result obtained in (3.8), we have that (3.9) is equal to:

$$(B_2 U_{2s} \square B_1 U_{1n} : B_2 U_{2n} \square B_1 U_{1s} : B_2 U_{2s} \square B_1 U_{1s})$$

And finally, the random effects matrix is

$$Z = (Z_2 \square X_1 : X_2 \square Z_1 : Z_2 \square Z_1), \quad (3.10)$$

where $Z_k = B_k U_{ks}$, for $k = 1, 2$.

■

As showed in the previous Chapter, we can consider a second order penalty and

define $\mathbf{X}_1 = [\mathbf{1}_n : \mathbf{x}_1]$, $\mathbf{X}_2 = [\mathbf{1}_n : \mathbf{x}_2]$. Then, we can expand the model matrices as

$$\mathbf{X} \equiv (\mathbf{1}_n : \mathbf{1}_n \square \mathbf{x}_1 : \mathbf{x}_2 \square \mathbf{1}_n : \mathbf{x}_2 \square \mathbf{x}_1), \quad (3.11)$$

$$\mathbf{Z} \equiv (\mathbf{Z}_2 \square \mathbf{1}_n : \mathbf{Z}_2 \square \mathbf{x}_1 : \mathbf{1}_n \square \mathbf{Z}_1 : \mathbf{x}_2 \square \mathbf{Z}_1 : \mathbf{Z}_2 \square \mathbf{Z}_1), \quad (3.12)$$

where by symbol “ \equiv ”, we denote that both matrices have the same elements but with a different ordering in the columns and block in the right-hand side of (3.11) and (3.12).

In the spatial context, this partition allows the representation of the fitted surface in terms of: a smooth term for the latitude effect, a smooth term for the longitude effect and a smooth term of the latitude and longitude interaction effect.

3.3 Examples of smoothing spatial data with P -splines

In this Section, we illustrate some examples of the smoothing mixed model methodology applied to spatial data. We consider the examples of spatial data shown in [Chapter 1](#) (i.e. geostatistical, regional data and point patterns). From an unified approach, given the set of spatial locations \mathbf{s} , we proceed as follows:

- For the n spatial locations, we construct the marginal B -spline bases, \mathbf{B}_1 and \mathbf{B}_2 , of dimensions $n \times c_1$ and $n \times c_2$, with ndx_1 and ndx_2 knots for each spatial dimension. We consider a cubic splines for the B -spline bases and a second order penalty. The spatial B -spline basis is constructed as $\mathbf{B}_2 \square \mathbf{B}_1$, of dimension $n \times c_1 c_2$.
- Reformulate the model into a mixed model using the reparameterization shown in Section 3.2, i.e.: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, with $\boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{G})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, with fixed and random effects matrices, \mathbf{X} and \mathbf{Z} defined in (3.3) and (3.4). And random effects covariance $\mathbf{G} = \sigma^2 \mathbf{F}$, where \mathbf{F} is the mixed model block-diagonal penalty in two-dimensions defined in (2.70).

$$\mathbf{F} = \begin{pmatrix} \lambda_2 \tilde{\boldsymbol{\Sigma}}_2 \otimes \mathbf{I}_{q_1} & & \\ & \lambda_1 \mathbf{I}_{q_2} \otimes \tilde{\boldsymbol{\Sigma}}_1 & \\ & & \lambda_1 \mathbf{I}_{c_2-q_2} \otimes \tilde{\boldsymbol{\Sigma}}_1 + \lambda_2 \tilde{\boldsymbol{\Sigma}}_2 \otimes \mathbf{I}_{c_1-q_1} \end{pmatrix}.$$

- Estimate the smoothing parameters λ_1 , λ_2 , and the variance of error term σ^2 , by maximization of the restricted log-likelihood (REML) function in (2.32).
- Estimate the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ from the standard mixed model equations in (2.30) and (2.31).

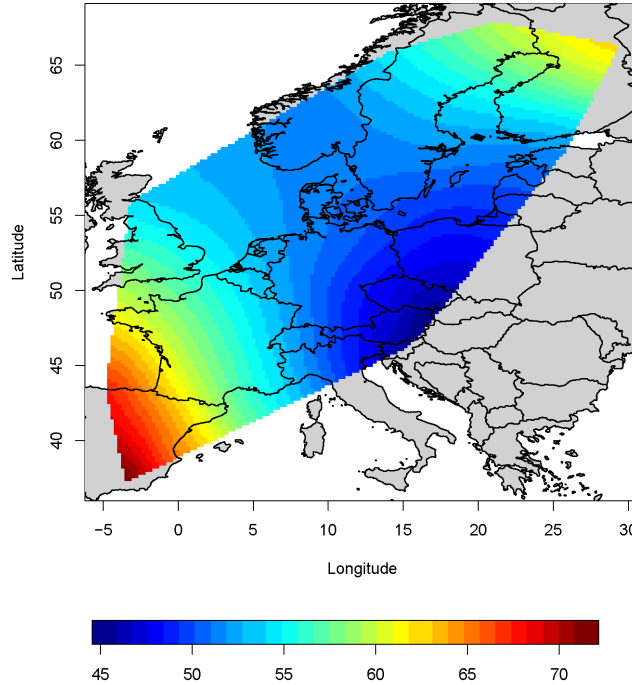


Figure 3.1: Smoothed surface of O_3 levels in January 1999.

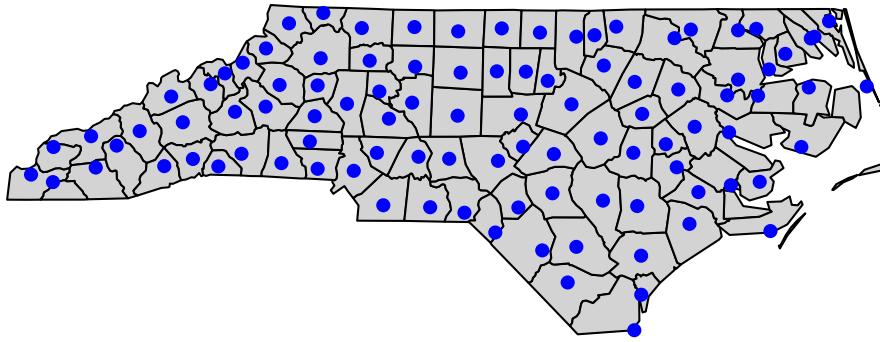
Smoothing geostatistical data with P -splines

As an example of geostatistical data, we consider the mean Ozone (O_3) levels in Europe in January 1999 shown in Section 1.1.1. We constructed the marginal B -spline bases with $ndx_1 = 10$ and $ndx_2 = 10$ knots. The estimated smoothing parameters were $\lambda_1 = 239.35$, and $\lambda_2 = 220.91$, and the error variance $\sigma^2 = 7.647$. Figure 3.1 illustrates the smoothed surface over the map of Europe. It reflects the spatial pattern of O_3 in January 1999, where the higher levels of O_3 in the south-west countries. We will study this data in the spatio-temporal setting in Chapter 4.

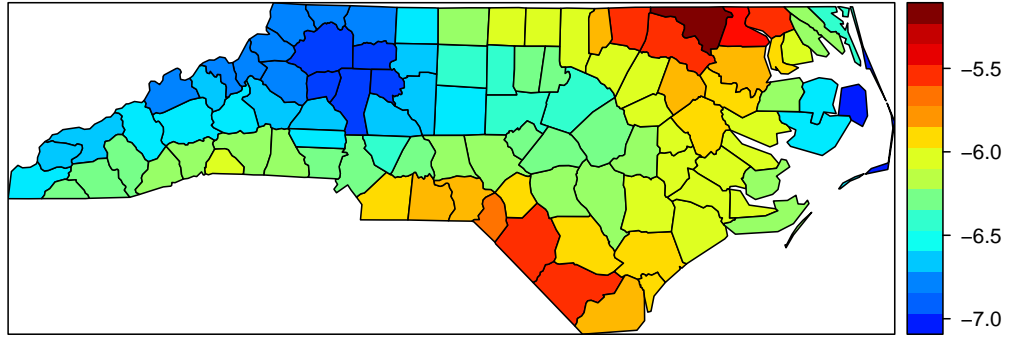
Spatial smoothing of regional data

For regional data, we consider as spatial locations s , the centroids of each of the regions. Then, for this type of data, we have $s = (x_1, x_2)$, where x_1 and x_2 , are the spatial coordinates of the centroids. Figure 3.2a shows the centroids of the 100 counties of North Carolina. The data set contains the number of counts of sudden-infant-deaths syndrome, so then we consider a P -GLMM as shown in Section 2.2.1. Thus, the observed number of SIDS counts, y is distributed as Poisson, $y \sim \text{Poisson}(\mu)$, where the

mean μ with log-link is $\mu = \exp(\eta)$, and η is the linear predictor modelled as a bivariate P -spline as a mixed model: $\eta = \mathbf{X}\beta + \mathbf{Z}\alpha$. The additional information of the number of births in 1974 was included as an offset term. The model was constructed with $ndx_1 = 15$ and $ndx_2 = 15$ knots, and was estimated by PQL as detailed in Section 2.2.2. The fitted spatial trend (i.e. $\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\alpha}$) is shown in Figure 3.2b, we obtained the values of the smoothing parameters $\lambda_1 = 6.67$ and $\lambda_2 = 130.65$, for longitude and latitude dimensions. The larger value of λ_2 indicates a smoother north-south effect for the SIDS data.



(a) Centroids of the 100 counties of North Carolina



(b) Spatial trend for SIDS data for year 1974.

Figure 3.2: Fitted smooth trend of SIDS data in 1974, using the centroids of the counties as spatial locations.

Spatial point pattern analysis

In the case of spatial analysis of spatial patterns, $s = (s_1, s_2)$, are the pair of spatial locations where the events of study have been located. For this type of spatial data, we have that the spatial domain \mathcal{D} is random, and therefore the events may occur in any location of the region of study. Thus, instead of constructing the spatial \mathbf{B} as $\mathbf{B}_2 \square \mathbf{B}_1$,

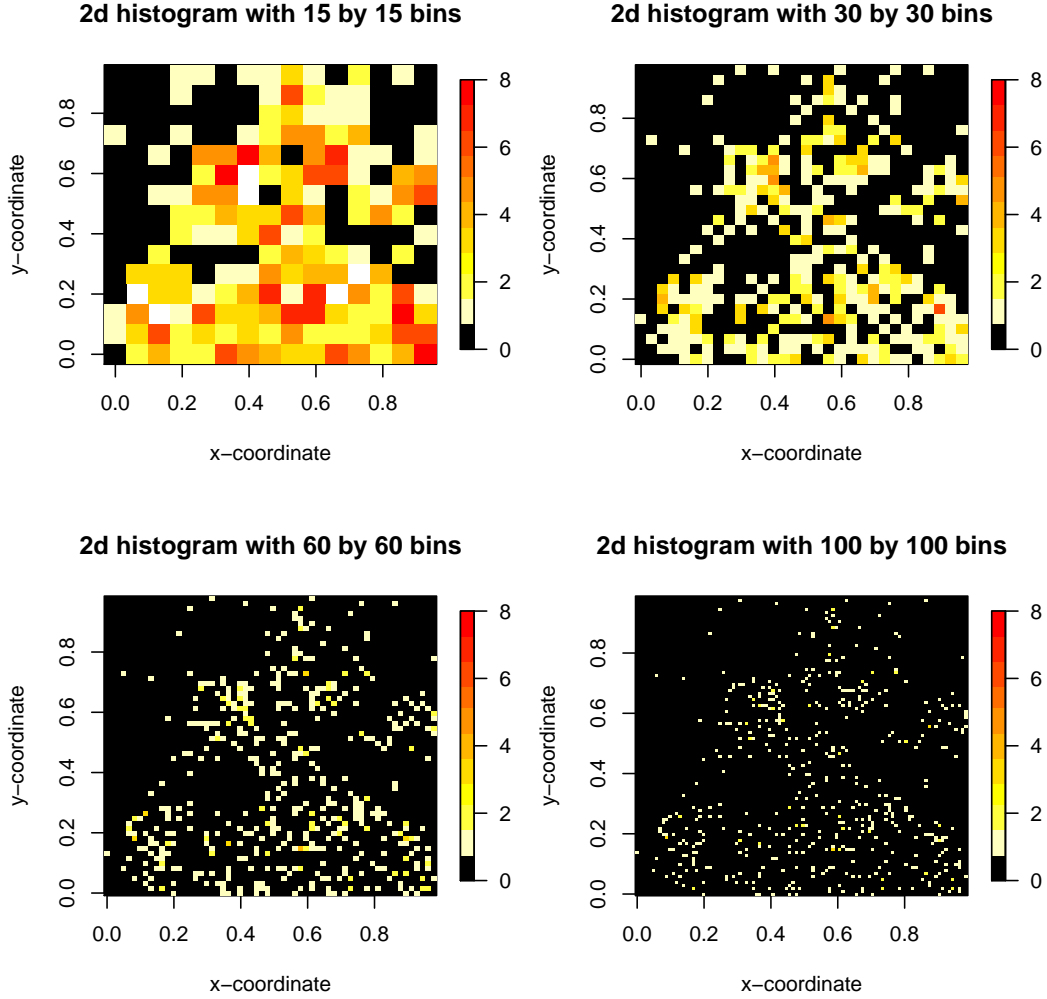


Figure 3.3: 2d histograms of counts of maples trees with different number of bins in each direction, $n_{bins} = \{15, 30, 60, 100\}$.

we can compute a two-dimensional histogram with $n_1 \times n_2$ bins and form an array \mathbf{Y} of Poisson counts in each bin, such that $\mathbf{y} = \text{vec}(\mathbf{Y})$, where \mathbf{x}_1 and \mathbf{x}_2 are the midpoints of the bins in each spatial dimension. The expected values can be arranged in the array \mathbf{M} , and then, the mean is $\boldsymbol{\mu} = \text{vec}(\mathbf{M})$. The two-dimensional P -spline Poisson model with log link and linear predictor is given by:

$$\boldsymbol{\eta} = \exp(\boldsymbol{\mu}) = \mathbf{B}\boldsymbol{\theta}, \quad (3.13)$$

where the regression basis \mathbf{B} is the Kronecker product of the marginal B -spline basis calculated from \mathbf{x}_1 and \mathbf{x}_2 , i.e. $\mathbf{B}_2 \otimes \mathbf{B}_1$, of dimensions $n_1 n_2 \times c_1 c_2$. This approach is

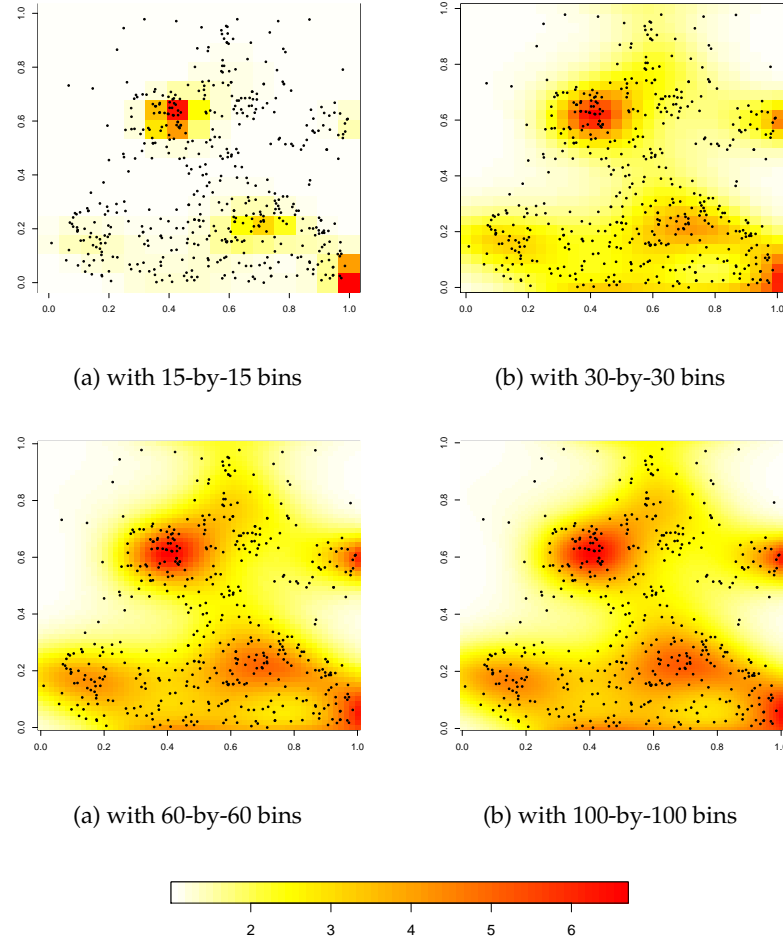


Figure 3.4: Smoothed intensity functions for different number of bins.

equivalent to consider a two-dimensional Poisson model for the intensity function, η , using a P -GLM approach. The main advantage is the possibility of using the GLAM methods [Currie et al. \(2006\)](#); [Eilers et al. \(2006\)](#), for fast and efficient computation. The vector of coefficients, θ can be arranged in a $c_1 \times c_2$ array Θ . Thus, (3.13) can be written as a GLAM: $\eta = \exp(M) = B_1 \Theta B_2'$. Using a mixed model formulation, the fixed and random effects matrices X and Z are those defined for data in array structure in (2.68) and (2.67). [Figure 3.3](#) shows the two-dimensional histograms of the maples trees in Langsing woods data set described in [Section 1.1.3](#) with different number of bins. It can be seen that as the number of bins increases, the resolution is finer. Given that, the number of bins and the number of knots in each direction (ndx_1 and ndx_2) are chosen by the analyst, the use of the GLAM approach provides a computationally fast exploratory tool to visualize spatial point patterns and estimate the intensity function. [Figure 3.4](#)

shows the smoothed intensities for different number of bins. The estimated smoothing parameters with 8 knots in the construction of the B -spline marginal bases of cubic splines and second order penalty, are shown in Table 3.2.

Table 3.2: Estimated smoothing parameters λ_1 and λ_2 for different number of bins.

Number of bins	Smoothing parameters	
	λ_1	λ_2
15×15	2.1305	28.8363
30×30	0.2258	4.9687
60×60	0.0435	0.9188
100×100	0.0126	0.2843

3.4 Smoothing mixed models for spatial count data

Areal data are very common in *disease mapping* applications, usually, this type of data are units such as counties, states or provinces, where the number of diseases counts are aggregated. Disease counts are assumed to be distributed as Poisson. From a P -spline approach, it is possible to consider the centroids of the areas as the geographical locations where the data are collected, and use a generalized linear mixed model (GLMM) approach and penalized quasi-likelihood (PQL) for estimation. We considered the use of the “Box-product” of the marginal B -spline basis and include individual area-effects as random effects to account for individual variation between regions. In Lee and Durbán (2009), we presented several alternatives to deal with overdispersion in spatial count data, one of them combines a smooth model (to account for the large-scale trend variability) with *conditional autoregressive* structured random effects (to account for the small-scale local variability) to yield a hybrid model called it *Smooth-CAR* model. The methodology is illustrated with the well known Scottish Lip Cancer data set.

3.4.1 Overdispersion in Poisson count data

Count data often presents overdispersion relative to a Poisson model. Overdispersion implies that the variance of the data is greater than the one expected under the distribution assumed. The assumption of Poisson distribution in count data has some limitations, since it assumes that the conditional mean and the conditional variance are both equal. Ignoring overdispersion may lead to serious problems in terms of underestimation of the standard errors and as a consequence for inference in the regression parameters. The problem of overdispersion has been considered from different approaches

in the literature (see for example [Breslow, 1984](#); [Lawless, 1987](#); [McCullagh and Nelder, 1989](#)). [Hinde and Demetrio \(1998\)](#) distinguish between two main approaches to deal with overdispersion:

- (i) assume a more general form of the variance function using additional parameters (for example with the incorporation of random effects), or
- (ii) consider more flexible distributional assumptions by letting the mean follow a Gamma distribution with mean μ and variance $\phi\mu$.

The first approach does not correspond to any specific distribution for the response variable, and can be considered as an extension of the basic Poisson model. The second approach is a mixture that leads to a Negative Binomial model. We consider both strategies in the P -GLMM framework as shown in Section 2.2.2. To illustrate the different approaches we consider the age-at-death data for greek females shown in Section 2.2.2.

The PRIDE approach

We start by considering the first alternative. [Perperoglou and Eilers \(2009\)](#) gave an approach based on P -splines, they use individual random effects which add extra parameters to the linear predictor of a Poisson GLM (with log link) for each observation; they used the acronym PRIDE (*Penalized Random Individual Dispersion Effects*) for this model. The PRIDE model is formulated as:

$$\eta = B\theta + \gamma I, \quad \gamma \sim \mathcal{N}(\mathbf{0}, \kappa^{-1} I), \quad (3.14)$$

where γ is an individual deviance effect vector (of length n) which provides a device to absorb the overdispersion which causes the extra variability. Note that the extra parameter γ is added to the linear predictor for each observation. Therefore, the model has more parameters than observations, and this yields an overparameterized model. However, it is possible to add a ridge penalty on γ (to shrink γ towards zero) to maintain the identifiability. [Perperoglou and Eilers \(2009\)](#) use an iterative algorithm in order to estimate the smooth curve ($B\theta$) and the extra parameter γ . We simplify their approach by the reparameterization of model (3.14) into a mixed model. Then, the linear predictor becomes:

$$\eta = X\beta + Z\alpha + \gamma I, \quad \alpha \sim \mathcal{N}(\mathbf{0}, G), \quad \gamma \sim \mathcal{N}(\mathbf{0}, \kappa^{-1} I), \quad (3.15)$$

Using PQL we obtain the following set of equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} & \mathbf{X}'\mathbf{W} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{G}^{-1} + \mathbf{Z}'\mathbf{W}\mathbf{Z} & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}\mathbf{X} & \mathbf{W}\mathbf{Z} & \kappa\mathbf{I} + \mathbf{W} \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \\ \gamma \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{z} \\ \mathbf{Z}'\mathbf{W}\mathbf{z} \\ \mathbf{W}\mathbf{z} \end{bmatrix}, \quad (3.16)$$

where \mathbf{z} is the *working vector*, $\mathbf{z} = \boldsymbol{\eta} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu})$, and \mathbf{W} is the diagonal matrix of weights, $\mathbf{W} = \text{diag}(\boldsymbol{\mu})$ and $\boldsymbol{\mu} = \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \gamma\mathbf{I})$. Equation (3.16) gives a very large system of equations, but it is possible to reduce it by defining γ as:

$$\gamma = \frac{\mathbf{W}}{\mathbf{W} + \kappa\mathbf{I}} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}). \quad (3.17)$$

If we define:

$$\mathbf{W}^* = \frac{\kappa\mathbf{W}}{\mathbf{W} + \kappa\mathbf{I}}. \quad (3.18)$$

Given (3.18), we have $\kappa\gamma = \mathbf{W}^*(\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha})$, using this result in (3.16), the system of equations (3.16) can be reduced to:

$$\begin{bmatrix} \mathbf{X}'\mathbf{W}^*\mathbf{X} & \mathbf{X}'\mathbf{W}^*\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}^*\mathbf{X} & \mathbf{G}^{-1} + \mathbf{Z}'\mathbf{W}^*\mathbf{Z} \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{W}^*\mathbf{z} \\ \mathbf{Z}'\mathbf{W}^*\mathbf{z} \end{bmatrix}. \quad (3.19)$$

This leads to the same set of equations as in a Poisson P -GLMM without overdispersion, but changing the matrix of weights to \mathbf{W}^* , and the addition of γ to the linear predictor. Then, the parameters $\hat{\beta}$ and $\hat{\alpha}$ are estimated as in (2.30) and (2.31), and $\hat{\gamma}$ as in (3.17). The covariance matrix \mathbf{V} is now given by:

$$\mathbf{V} = \mathbf{W}^{*-1} + \mathbf{Z}\mathbf{G}\mathbf{Z}'. \quad (3.20)$$

Then, conditional on the estimates of the fixed and random effects, we estimate the smoothing parameter λ and the dispersion effect κ by REML as in Equation (2.32). The PQL solution is obtained by iteration until convergence.

P -GLMM for Negative Binomial model

The second alternative to model count data in the presence of overdispersion is the specification of a distribution that permits more flexible modelling of the variance than the Poisson distribution. The standard parametric model to account for overdispersion is the Negative Binomial. The most common way to derive this distribution is through a Poisson-Gamma mixture. This is a two-stage model that assumes that data are Poisson, but there is a heterogeneity that it is not observed. The Negative Binomial has been

derived and presented with different reparameterizations (see [Cameron and Trivedi, 1998](#), Chapter 3). We follow the one derived by letting the mean of the Poisson distribution vary according to a parameter ζ given by the Gamma distribution. The stochastic component is given by

$$\begin{aligned} \mathbf{y}|\zeta &\sim \text{Poisson}(\zeta\mu), \quad \text{and} \\ \zeta &\sim \frac{1}{\kappa}\text{Gamma}(\kappa). \end{aligned}$$

The marginal distribution of \mathbf{y} is, then, the Negative Binomial with mean μ and variance $\mu + \mu^2/\kappa$, where κ is the dispersion parameter. Note that, for a large value of κ , the Negative Binomial model reduces to Poisson. Then, we have the response \mathbf{y} defined as:

$$\mathbf{y} \sim \text{Neg Bin}(\mu, \kappa)$$

with density function:

$$\begin{aligned} f(\mathbf{y} = y_i|\mu, \kappa) &= \frac{\Gamma(y_i + \kappa)}{\Gamma(y_i + 1)\Gamma(\kappa)} \left(\frac{\kappa}{\kappa + \mu_i} \right)^\kappa \left(\frac{\mu_i}{\kappa + \mu_i} \right)^{y_i}, \\ &\text{for } \kappa \geq 0, \text{ and } y_i = 0, 1, 2, \dots \end{aligned}$$

and probability function:

$$P(\mathbf{y} = y_i|\mu_i, \kappa) = \binom{y_i + \kappa - 1}{y_i} \left(\frac{\mu_i}{\kappa + \mu_i} \right)^{y_i} \left(\frac{\kappa}{\kappa + \mu_i} \right)^\kappa, \quad (3.21)$$

The fact that the Negative Binomial distribution has two variance parameters and is not in the exponential family, makes more difficult the extension of the methodology developed for Poisson data. However, it can be formulated as a GLM, if the parameter κ is assumed constant (see [Thurston et al., 2000](#)). Thus, the log-likelihood is

$$\begin{aligned} \mathcal{L}(\mu_i, \kappa|y_i) &= y_i \ln \left(\frac{\mu_i}{\mu_i + \kappa} \right) - \kappa \ln \left(\frac{\mu_i}{\mu_i + \kappa} \right) \\ &\quad + \ln \Gamma(y_i + \kappa) - \ln \Gamma(\kappa) - \ln \Gamma(y_i + 1) + \kappa \ln \kappa, \end{aligned} \quad (3.22)$$

from which we can see that the canonical link is $\eta_i = \ln(\mu_i/\mu_i + \kappa)$.

If κ were known, this would be an exponential family. For a given κ , the log-likelihood for the vector $\boldsymbol{\mu}$ is

$$\mathcal{L}(\boldsymbol{\mu}; \kappa) = \sum_{i=1}^n y_i \ln \left\{ \frac{\mu_i}{(\mu_i + \kappa)} \right\} - \sum_{i=1}^n \kappa \ln \left(\frac{1 + \mu_i}{\kappa} \right) + c(y, \kappa),$$

Table 3.3: Comparison of smoothing mixed models for count data.

Model	log link	Inverse link	Weight matrix
Poisson	$\eta = \mathbf{X}\beta + \mathbf{Z}\alpha$	$\mu = e^\eta$	$\mathbf{W} = \text{diag}(\mu)$
PRIDE	$\eta = \mathbf{X}\beta + \mathbf{Z}\alpha + \gamma\mathbf{I}$	$\mu = e^\eta$	$\mathbf{W}^* = \frac{\kappa \text{diag}(\mu)}{\text{diag}(\mu) + \kappa\mathbf{I}}$
Neg. Binomial	$\eta = \mathbf{X}\beta + \mathbf{Z}\alpha$	$\mu = e^\eta$	$\mathbf{W} = \kappa \text{diag}(\frac{\mu}{\kappa + \mu})$

where $c(y, \kappa)$ is a function of the y_i 's and κ . For a given μ , the log likelihood for κ is

$$\begin{aligned} \mathcal{L}(\mu_i, \kappa) &= n \{ \kappa \ln \kappa - \ln \Gamma(\kappa) \} + \\ &+ \sum_{i=1}^n \{ \ln \Gamma(y_i, \kappa) - (y_i + \kappa) \ln(\kappa + \mu_i) \} + d(y, \mu) \end{aligned} \quad (3.23)$$

for some function $d(y_i, \kappa)$. [Thurston et al. \(2000\)](#) suggested the use of the log-link to overcome some difficulties with the canonical link. Then, we can obtain the GLM-based Negative Binomial that yields identical parameters estimates to the Poisson-Gamma mixture.

We can derive the penalized likelihood as in the Poisson GLMM case shown in Section 2.2.2, with joint density:

$$\begin{aligned} f(\mathbf{y}|\alpha) &= \exp \left[\mathbf{y}'(\mathbf{X}\beta + \mathbf{Z}\alpha - \log \{ \kappa \mathbf{1} + \exp(\mathbf{X}\beta + \mathbf{Z}\alpha) \}) \right. \\ &\quad \left. - \kappa \mathbf{1}' \log \{ \kappa \mathbf{1} + \exp(\mathbf{X}\beta + \mathbf{Z}\alpha) \} \right] \\ &\quad + \exp \left[n\kappa \log(\kappa) + \mathbf{1}' \log(\mathbf{y} + \kappa \mathbf{1}) - n \log(\Gamma(\kappa)) \right], \end{aligned} \quad (3.24)$$

And penalized log-likelihood

$$\mathcal{L}_p(\beta, \alpha, \kappa, \sigma_\alpha^2) = \mathcal{L}(\beta, \alpha, \kappa, \sigma_\alpha^2) - \frac{1}{2} \alpha' \mathbf{G}^{-1} \alpha, \quad (3.25)$$

where $\mathcal{L}(\cdot)$ is the ordinary log-likelihood, and \mathbf{G} is the covariance matrix for the random effects. In the smoothing mixed model approach, the matrix \mathbf{G} depends on the smoothing parameters and a block-diagonal matrix \mathbf{F} . Taking the derivative of (3.25) with respect to β and α , yields:

$$\mathbf{X}' \left(\frac{\mathbf{y} - \exp(\mathbf{X}\beta + \mathbf{Z}\alpha)}{\kappa \mathbf{1} + \exp(\mathbf{X}\beta + \mathbf{Z}\alpha)} \right) = 0 \quad (3.26)$$

$$\kappa \mathbf{Z}' \left(\frac{\mathbf{y} - \exp(\mathbf{X}\beta + \mathbf{Z}\alpha)}{\kappa \mathbf{1} + \exp(\mathbf{X}\beta + \mathbf{Z}\alpha)} \right) = \mathbf{G}^{-1} \alpha. \quad (3.27)$$

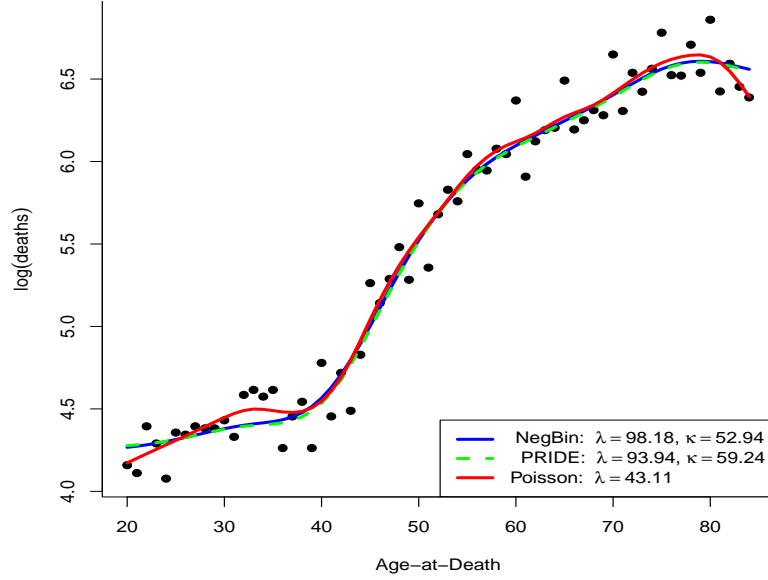


Figure 3.5: Comparison of fitted curves for smoothing mixed models: Poisson, PRIDE and Negative Binomial

Equations (3.26) and (3.27) are similar to the score equations in the Poisson case, but with matrix of weights given by:

$$\mathbf{W} = \kappa \operatorname{diag} \left(\frac{\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})}{\kappa \mathbf{1} + \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})} \right). \quad (3.28)$$

The estimation of the parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and the variance components $(\sigma_{\alpha}^2, \kappa)$ is done iteratively by PQL as shown for the case of Poisson data. Note that the weight matrix (3.28) of the Negative Binomial GLMM is similar to the one in the PRIDE model, \mathbf{W}^* in (3.18). The difference is that \mathbf{W}^* includes the parameter γ in the linear predictor. Table 3.3 shows the main differences in count data regression models proposed. We can also estimate the hat-matrix and confidence intervals as shown in Chapter 2. For model selection criteria AIC and BIC can be computed using the deviance for known κ defined as:

$$\operatorname{Dev}_{nb} = \sum_{i=1}^n \left\{ y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i + \kappa) \ln \left[\frac{y_i + \kappa}{\hat{\mu}_i + \kappa} \right] \right\}, \quad (3.29)$$

(See Cameron and Trivedi, 1998, pg. 153).

In order to illustrate the performance of the smoothing mixed models for overdispersed count data, we considered the age-at-death greek females mortality data studied

by [Kostaki and Panousis \(2001\)](#). We already used this data in Section 2.1.4, to present the P -GLMM for Poisson data.

We consider 15 knots to construct the model B -splines basis B , with a cubic spline a second order penalty. We reparameterize the model into a mixed model formulation and fit Poisson, PRIDE and Negative Binomial smooth mixed models as summarized in Table 3.3, using REML and PQL for estimation. Figure 3.5 shows the fitted curves for the three models. The PRIDE and Negative Binomial smooth mixed models allow to estimate the overdispersion in Poisson counts, with values of overdispersion parameter κ equal to 52.94 and 59.24, respectively. These models also present a similar smooth fitted curves (with similar values of the smoothing parameter λ), and greater than the smoothing parameter for the Poisson fit (that is not able to capture the extravariability). This also reflects that PRIDE and Negative Binomial fitted curves are smoother than Poisson. These results are not very surprising. In the simulation study presented by [Perperoglou and Eilers \(2009\)](#), they showed that PRIDE model performs as well as (and in some cases better) than the Negative Binomial, even when the true model is the Negative Binomial. For the case of overdispersed count data, in fact, the weights are the same as suggested by [Thurston et al. \(2000\)](#), thus, the value of κ from PRIDE model and Negative Binomial should be similar. Then PRIDE model can be considered as an approximation of the Negative Binomial distribution. Moreover, the PRIDE model allow us to consider the underlying stochastic process in terms of a smooth trend with an unexplained heterogeneity as overdispersion.

3.4.2 Spatial smoothing mixed models with CAR structure

The most popular approach in modelling spatial dependency structure for lattice or regional data are the conditionally autoregressive (CAR) models introduced by [Besag \(1974\)](#). These hierarchical models allow both spatially structured variability and unstructured heterogeneity by assuming a prior distribution for the spatial effects considering the neighboring regions. These models have been widely used in the context of regional data ([Besag and Green \(1993\)](#); [Leroux et al. \(1999\)](#); [Dean et al. \(2001\)](#); [Congdon \(2006\)](#); [Wakefield \(2007\)](#) among others). *Neighborhoods* can be defined by several criteria, depending on the shape of the lattice, for example, the distance between the centroids of the regions, bordering regions or sharing a common border with a given region, (see [Cressie and Chan, 1989](#); [Besag et al., 1991](#); [Besag and Kooperberg, 1995](#)). However, when applying these CAR models to irregular lattices, the imposed neighborhood structure and the spatial correlation could be misleading and strongly dependent to the number of neighbors. Furthermore the neighborhood criteria must be sometimes carefully examined. For instance, in the case of very irregular regions with different sizes

and shapes or in the presence of not contiguous regions like islands.

In Lee and Durbán (2009), we proposed the use of the smoothing mixed model approach for spatial count data given in previous section, together with a conditionally autoregressive (CAR) structure. We call these models “Smooth-CAR” models. The smooth component let us model the spatial trend along larger geographical distances, and the local (non-smooth) correlation is taken into account by means of a CAR component. We intend to separate the global trend and the purely individual regional effect. The mixed model representation of P -splines allows us to fit the model as a GLMM. This is a very challenging task since it is not clear whether and when both effects are identifiable and further research still needs to be done, but as we will see in the next section, in the example analyzed, this method performs better than the traditional spatial models and it gives a clearer picture of the spatial variation in the data.

The model proposed is:

$$\eta = \underbrace{X\beta + Z\alpha}_{f(x_1, x_2)} + \underbrace{\mathbf{b}}_{\text{CAR}}, \quad (3.30)$$

where $X\beta + Z\alpha$ corresponds to the mixed model representation of the bivariate spatial P -spline showed in Section 3.2. Now the random effect $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_b)$, has covariance matrix given by a CAR model.

The basic spatial (intrinsic) CAR model (Besag et al., 1991) uses the adjacencies to define neighborhoods in a conditional specification of the model. This model considers the covariance matrix \mathbf{G}_b as a sum of two separate variance components to represent both spatial and non-spatial correlation. One component models the spatially-structured variation, and another models the unstructured or individual region-level heterogeneity in the data. In this case, \mathbf{G}_b has the form

$$\mathbf{G}_b = \sigma_s^2 \mathbf{Q}^- + \kappa^{-1} \mathbf{I} \quad (3.31)$$

where $\mathbf{Q} = \{q_{i,j}\}$ is a $n \times n$ matrix determined by the neighborhood structure of the regions. And \mathbf{Q}^- denotes the Moore-Penrose Generalized inverse of matrix \mathbf{Q} . The i^{th} diagonal elements of \mathbf{Q} are the number of neighbors in the i^{th} region. The elements out of the diagonal are

$$q_{i,j} = \begin{cases} -1 & \text{if } i^{th} \text{ and } j^{th} \text{ regions are neighbors} \\ 0 & \text{otherwise.} \end{cases}$$

Alternative formulations of Besag’s model (3.31) have been proposed in the literature. For example, Leroux et al. (1999) adopts a prior specification of the random effects

\mathbf{b} with covariance matrix \mathbf{G}_b given by:

$$\mathbf{G}_b = \sigma_s^2 (\phi \mathbf{Q} + (1 - \phi) \mathbf{I})^{-1}, \quad (3.32)$$

where σ_s^2 is variance of the random effects and ϕ is a spatial autocorrelation parameter. Dean et al. (2001) assumes a different formulation of \mathbf{G}_b :

$$\mathbf{G}_b = \sigma_s^2 (\phi \mathbf{Q}^- + (1 - \phi) \mathbf{I}). \quad (3.33)$$

Note that (3.33) is a reparameterization of the model (3.31). In both models, (3.32) and (3.33) the covariance parameters are identifiable and ϕ , measures the relative weight between *structured* and *unstructured* variability ($0 \leq \phi \leq 1$), when $\phi = 1$, all the overdispersion is due to the spatial correlation so there is no unstructured heterogeneity and model (3.33) is equivalent to the intrinsic CAR model in (3.31). When $\phi = 0$, there is an absence of spatial correlation in the data and the overdispersion is not caused by a spatial heterogeneity, and the model reduces to the PRIDE model in (3.15), with $\sigma_s^2 = \kappa^{-1}$. If $0 < \phi < 1$, the random effects are correlated and the data presents a combination of spatial structured and unstructured component.

The Smooth-CAR model (3.30), can be reformulated as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^* \mathbf{u}, \quad \text{with} \quad \mathbf{Z}^* = [\mathbf{Z} : \mathbf{I}] \quad (3.34)$$

and the random effect $\mathbf{u} = (\boldsymbol{\alpha}, \mathbf{b})'$, has a block-diagonal covariance matrix:

$$\mathbf{G}_u = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_b \end{pmatrix}, \quad (3.35)$$

where \mathbf{G} is the covariance matrix for the random effects $\boldsymbol{\alpha}$ of the bivariate P -spline (which depends on the smoothing parameters λ_1, λ_2 and the block-diagonal penalty \mathbf{F} as in (2.70)), and \mathbf{G}_b is the covariance matrix of the CAR random effect which depends on the parameters κ and σ_b^2 or ϕ depending on the type of CAR model considered. The estimation of the model parameters can also be done by PQL and REML. The matrix \mathbf{V} becomes now $\mathbf{W}^{-1} + \mathbf{Z}^* \mathbf{G}_u \mathbf{Z}^{*'} and the vector of random effects \mathbf{u} is estimated as$

$$\hat{\mathbf{u}} = \mathbf{G}_u \mathbf{Z}^{*'} \mathbf{V}^{-1} (\mathbf{z} - \mathbf{X} \hat{\boldsymbol{\beta}}). \quad (3.36)$$

Application to Scottish Lip Cancer data

We illustrate the methodology proposed above with the analysis of the Scottish Lip cancer data. The data set consists on the observed (y) and expected (e) number of cases of lip cancer registered in 56 counties in Scotland during the period 1975-1980. This data set has been analyzed several times in the literature (see Wakefield (2007) for a detailed review). Clayton and Kaldor (1987) analyzed the observed and expected counts using Empirical Bayes estimation, and used several alternatives for the distribution of the random effects. Breslow and Clayton (1993) proposed a conditional independent Poisson model, where the random effect is modelled by Gaussian intrinsic autoregression. A different approach is taken by Yasui and Lele (1997), the hierarchical modelling for spatial disease rates is based on estimating functions. This method led to simpler computations as in the P -splines case, both approaches are very attractive when data sets are large. The models presented by Dean et al. (2001) and Militino et al. (2001) are a reparametrization of Besag (1984), and allow for the determination of the relative weights between spatial and unstructured variation (these models have already been presented in the previous section). Finally, in the last few years, Congdon (2006) used a generalized additive form that allows regression to vary over regions, and Congdon (2007) considered continuous and discrete priors that account for risks that are discordant with those of neighboring areas. We fitted several models to this data set:

(i) Smooth P -spline models

- $\eta = \log(e) + f(\text{lon}, \text{lat})$, where $\log(e)$ is the offset term (**Poisson** model)
- $\eta = \log(e) + f(\text{lon}, \text{lat}) + \gamma I$, $\gamma \sim \mathcal{N}(\mathbf{0}, \kappa^{-1} I)$, (**PRIDE** model), and
- The **Negative Binomial** version of the Poisson model presented in Section 3.4.1.

(ii) Hierarchical CAR models: $\eta = \log(e) + \mathbf{X}\beta + \mathbf{b}$, $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_b)$, with

$$\mathbf{G}_b = \sigma_s^2 \mathbf{Q}^- + \kappa^{-1} I \quad (\text{Besag model})$$

$$\mathbf{G}_b = \sigma_s^2 (\phi \mathbf{Q}^- + (1 - \phi) I) \quad (\text{Dean model})$$

$$\mathbf{G}_b = \sigma_s^2 (\phi \mathbf{Q} + (1 - \phi) I)^{-1} \quad (\text{Leroux model})$$

In order to compare the proposed models we use the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC). For the CAR models, we follow an Empirical instead of fully Bayes approach (Clayton and Kaldor, 1987) in order to ease model comparisons. It should be noticed that in the case of the Negative Binomial model, the adequacy of the fitted model must be carefully considered since we

Table 3.4: Comparisons of fitted models to Scottish Lip Cancer data.

Model		Parameters					AIC	BIC	ED
		λ_1	λ_2	σ_s^2	κ^{-1}	ϕ			
Smooth:	<i>Poisson</i>	11.75	3.63	-	-	-	114.04	228.46	15.90
	<i>PRIDE</i>	30.12	5.50	-	0.12	-	89.64	180.08	31.73
	<i>PRIDE*</i>	10.45		-	0.12	-	89.82	180.46	32.31
	<i>Neg. Bin</i>	8.45	1.34	-	0.10	-	72.63	145.56	11.91
CAR:	<i>Besag</i>	-	-	0.78	10^{-6}	-	89.36	179.56	32.78
	<i>Dean</i>	-	-	0.78	-	0.99	89.36	179.56	32.78
	<i>Leroux</i>	-	-	0.78	-	0.99	89.36	179.56	32.78
Smooth-CAR:	<i>Besag</i>	30.40	18.28	0.55	-	-	87.48	175.75	30.67
	<i>Dean</i>	30.37	18.21	0.55	-	0.99	87.49	175.77	30.67
	<i>Leroux</i>	30.11	16.37	0.53	-	0.97	87.46	175.70	30.64

are assuming different distribution for the data. For the Negative Binomial we use the definition of the deviance in (3.29).

The results obtained are summarized in Table 3.4. The smooth surface in the Poisson model is fitted using two-dimensional P -splines, where the B -splines basis was constructed from marginal basis, the number of knots was 15 for each basis and the penalties had order two. The PRIDE model incorporates the spatial random effects (γ) for each of the 56 counties which allowed us to consider individual characteristics of each county and the possible unstructured variation. The estimation of the spatial effects resulted in the higher values of the effective dimension of the PRIDE model respect to Negative Binomial and Poisson models. Figure 3.8 illustrates the smooth large-scale spatial trend of PRIDE model and the unstructured variation between counties. It can be seen clearly an increasing trend from the more central counties to the ones on the coast, and also from south to north. We have also fitted an isotropic version (both smoothing parameters are equal) of the PRIDE model (See PRIDE* in Table 3.4), the AIC criteria is slightly lower for the anisotropic model, although there is not much difference. However, it is possible that, even in the situation where both covariates (longitude and latitude) are measured in the same scale, using a single smoothing parameter might not be the appropriate choice.

Figure 3.6 shows two different adjacency matrix for the Scottish data set. Considering the common border criteria, the isles of Shetland, Orkney and Western Isles have no neighbors and the total number of neighbors is 234. We used the adjacency matrix defined by Breslow and Clayton (1993) in order to fit the CAR models with a number of 264 neighbors. We show the results obtained with this last matrix, since it is the one

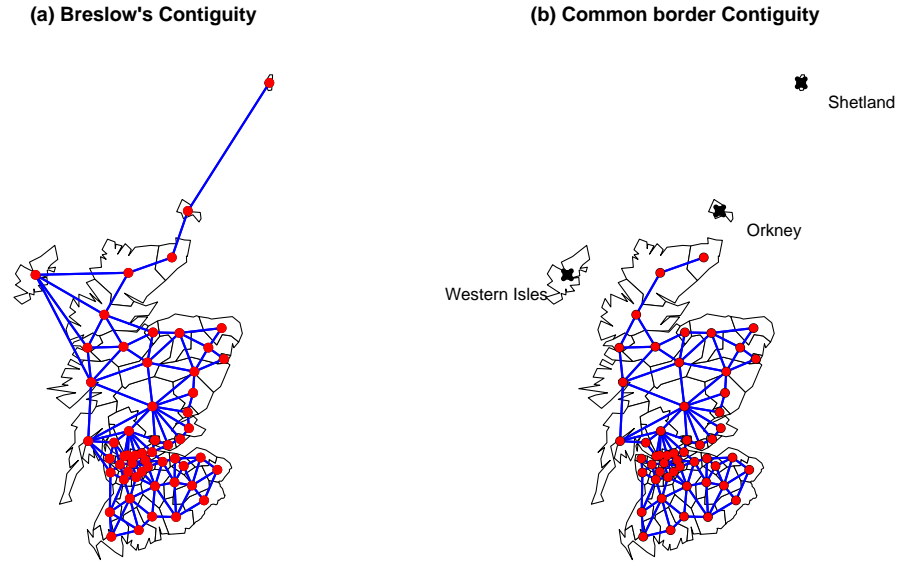


Figure 3.6: Neighboring structure for Scottish data: (a) Contiguity defined in [Breslow and Clayton \(1993\)](#); (b) contiguity based on sharing a common border.

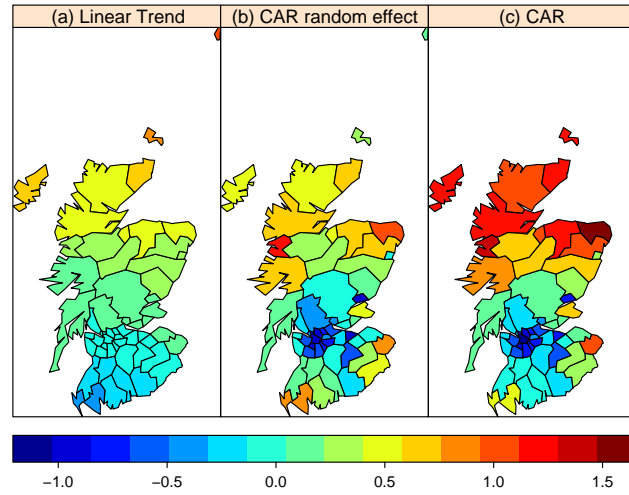


Figure 3.7: CAR model: (a) Linear Trend ($X\beta$); (b) CAR random effect (b) with G_b defined by Dean in (3.33) and (c) CAR model fit ($X\beta + b$).

commonly used in the literature. However, it is worth mentioning that results were different depending on the neighborhood criteria used.

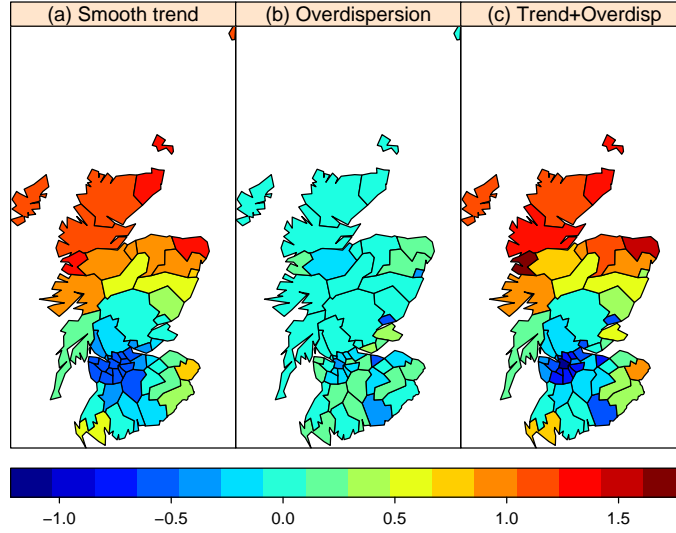


Figure 3.8: PRIDE Model: (a) Spatial Smooth Trend ($X\beta + Z\alpha$); (b) Overdispersion individual random effects (γ) and (c) the sum of trend and overdispersion effects.

For the CAR models (see Figure 3.7), the parameter ϕ presents high values (≈ 1) for both models (3.32) and (3.33), which denotes that all the variation is explained by the spatial autocorrelation. For the intrinsic Besag's CAR model, similar interpretation can be obtained for the estimated parameters, the spatial correlation structure absorbs the overdispersion without variability in each region, and the three alternative CAR formulations present similar results on model parameters.

Table 3.4 shows the better performance of Smooth-CAR models in terms of AIC and BIC criteria. As we mentioned above, it is important to noticed that, although the AIC and BIC values of the Negative Binomial are smaller than the ones obtained in other models, they cannot be compared, since the distribution assumed for the data is different. In Figure 3.9 we can see both, large geographical trend and local spatial variation. If we compare Figure 3.9 with Figure 3.8, we can see that in the Smooth-CAR model the large-scale trend is smoother than in the PRIDE model. This could be expected since in the PRIDE model all the spatial variation is fitted by the P -spline. The partition of the spatial variation seems more realistic in Figure 3.9. However, more research is still needed to check to what extent it is possible to separate both spatial effects, of whether we could really only look at the overall fit.

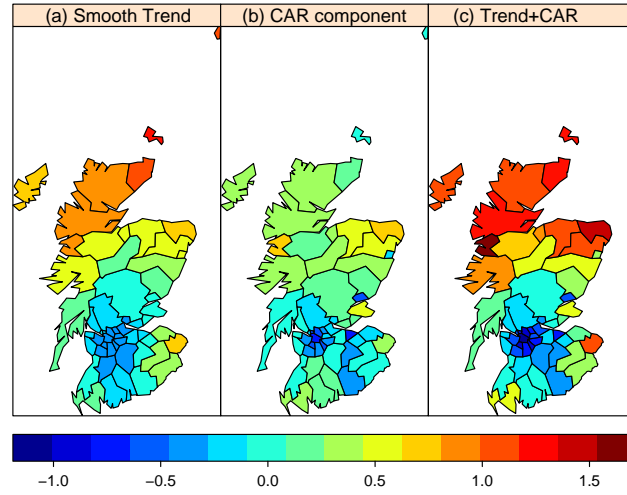


Figure 3.9: Smooth-CAR model: (a) Smooth Trend ($X\beta + Z\alpha$); (b) CAR structured random effects (b) and (c) the sum of trend and CAR component.

We also performed a residual analysis. We consider the deviance residuals for Poisson GLM data, defined as:

$$r_d = \text{sign}(y - \hat{\mu}) \sqrt{2\{y \log(y/\hat{\mu}) - y + \hat{\mu}\}},$$

(see [Cameron and Trivedi, 1998](#), pg. 141). Figure 3.10a show the map plots of the deviance residuals for Poisson, PRIDE, Smooth-CAR and CAR models (with Dean's covariance structure). The deviance residuals for the Poisson model reflects the variability not captured by the spatial Poisson P -spline. For the rest of the models, the spatial deviance residuals exhibit a small-scale spatial dependence, located around the high populated urban areas like Glasgow, Dundee or Edinburgh or large regions as Inverness and Annandale (see Figure 3.10b).

3.5 Simulation study

In order to compare the different P -spline models for spatial count data presented in [Section 3.4](#), we conducted a simulation study with different settings. We considered simulations of count data over the map of Scotland. As spatial locations we considered the centroids of the 56 districts levels in the map of Scotland. The aim of the study is

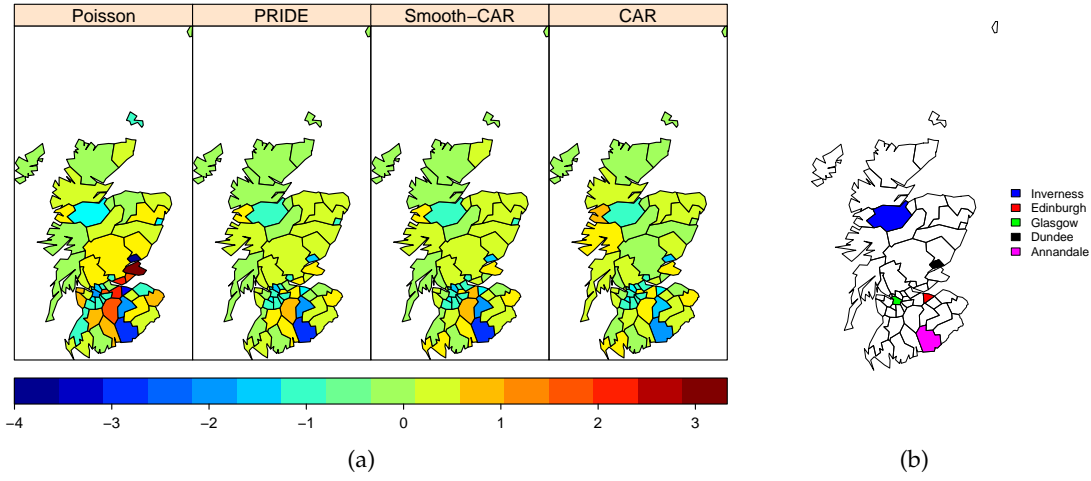


Figure 3.10: (a) Spatial deviance residuals for fitted models and (b) locations of regions of Scotland with larger values of the residuals.

to check the performance of spatial PRIDE, and Smooth-CAR models, with respect to spatial Poisson and CAR models. We simulated Poisson count data of the number of lip cancer cases in four different scenarios, with mean μ , and log link: $\log(\mu) = \log(e) + \eta$, where η is the linear predictor, and e is the expected number of counts of lip cancer from the original data, each of them depends on the simulated predictor η . Now, we summarize the simulated scenarios and the fitted models for each of them:

- **Scenarios 1 and 2**, we simulated counts from:

$$\begin{aligned}\eta^{(1)} &= f_{\text{non-linear}}, & (“\text{non-linear trend}”) \\ \eta^{(2)} &= f_{\text{non-linear}} + \gamma, & (“\text{non-linear trend with overdispersion}”) \end{aligned}$$

where $f_{\text{non-linear}}$ is a smooth non-linear trend, and $\gamma \sim \mathcal{N}(0, \kappa^{-1} \mathbf{I})$. For these two scenarios, we fitted the following models:

- (i) Poisson P -spline model:

$$- \eta = \log(e) + \mathbf{X}\beta + \mathbf{Z}\alpha.$$

- (ii) PRIDE P -spline model:

$$- \eta = \log(e) + \mathbf{X}\beta + \mathbf{Z}\alpha + \gamma \mathbf{I}, \text{ with } \gamma \sim \mathcal{N}(0, \kappa^{-1} \mathbf{I}).$$

- (iii) Smooth-CAR model:

$$- \eta = \log(e) + \mathbf{X}\beta + \mathbf{Z}\alpha + \mathbf{b}, \text{ with } \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_b).$$

- (iv) CAR model:

$$- \boldsymbol{\eta} = \log(\mathbf{e}) + \mathbf{X}\boldsymbol{\beta} + \mathbf{b}, \text{ with } \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_b).$$

Since we are considering as true models a non-linear trend with and without overdispersion, for models (iii) and (iv) we considered a covariance for the CAR random effect \mathbf{b} , defined as: $\mathbf{G}_b = \sigma_s^2 \mathbf{Q}^- + \kappa^{-1} \mathbf{I}$ (Besag's model).

- **Scenarios 3 and 4**, we simulated counts from:

$$\begin{aligned} \boldsymbol{\eta}^{(3)} &= f_{\text{non-linear}} + \mathbf{b}, & (\text{"non-linear trend with CAR random effect"}) \\ \boldsymbol{\eta}^{(4)} &= f_{\text{linear}} + \mathbf{b}, & (\text{"linear trend with with CAR random effect"}) \end{aligned}$$

where \mathbf{b} is a vector simulated from a CAR Normal distribution with zero mean, and covariance matrix as defined by Dean's model, i.e.: $\mathbf{G}_b = \sigma_s^2(\phi \mathbf{Q}^- + (1 - \phi) \mathbf{I})$, since we are interested in evaluate the performance of the fitted model with different combinations of the structured variability (σ_s^2) and the un-structured variability, controlled by the parameter $\phi \in (0, 1)$. For both scenarios 3 and 4, we fitted Poisson and PRIDE P -spline models as in previous scenarios and Smooth-CAR and CAR models with \mathbf{G}_b as defined by Dean's model.

The spatial linear/non-linear trends were selected as the fitted CAR and PRIDE trends to the original Scottish lip cancer data, shown in Section 3.4.2 with results summarized in Table 3.4, i.e.:

$$\begin{aligned} f_{\text{linear}} &= \log(\mathbf{e}) + \mathbf{X}\hat{\boldsymbol{\beta}}, \text{ and} & (\text{"CAR trend"}) \\ f_{\text{non-linear}} &= \log(\mathbf{e}) + \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\boldsymbol{\alpha}}. & (\text{"PRIDE trend"}) \end{aligned}$$

These spatial trends were shown in Figure 3.7(a), and Figure 3.8(a), respectively.

We simulated $R = 100$ data sets, and in order to compare the models performance, for each fitted model, we computed the logarithm of the empirical Mean Square Error (MSE) given by:

$$\text{MSE}(\hat{\boldsymbol{\eta}}_i^{(r)}) = \frac{1}{56} \sum_{s=1}^{56} \left(\boldsymbol{\eta}_i - \hat{\boldsymbol{\eta}}_i^{(r)} \right)^2, \text{ for } r = 1, \dots, R \quad (3.37)$$

where $\boldsymbol{\eta}_i$, denotes the true linear predictor for the i^{th} scenario, and $\boldsymbol{\eta}_i^{(r)}$, the estimated linear predictor for each fitted model. Smaller values of MSE indicates accurate results of the selected model in capturing the true linear predictor.

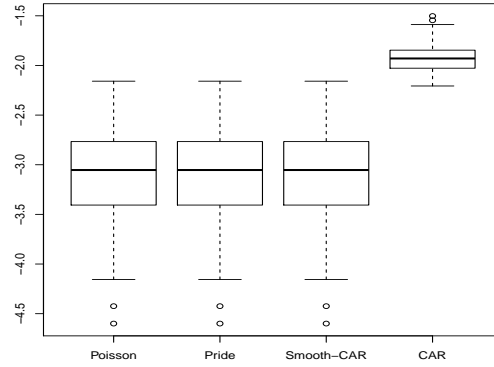


Figure 3.11: $\log(\text{MSE})$ comparison of Poisson, PRIDE, Smooth-CAR and CAR models in scenario 2 with $R = 100$.

Scenario 1: Smooth spatial non-linear trend

In this case, the true model consists in a spatial non-linear trend, then a Poisson P -spline fit will correspond to the most adequate model. In this scenario, we are interested in analyzing the performance of the alternative smooth models (PRIDE and Smooth-CAR) with respect to the CAR model.

Table 3.5: Mean and standard errors of estimated parameters for Scenario 1.

Model		$\hat{\sigma}_s$	$\hat{\kappa}^{-1}$
Smooth	PRIDE	-	$< 10^{-5}$
		-	$(< 10^{-3})$
	Smooth-CAR*	$< 10^{-5}$	$< 10^{-5}$
		-	$(< 10^{-3})$
CAR*		0.5591	10^{-3}
		(0.0818)	(0.0062)

* Models fitted with Besag's model CAR covariance structure.

Table 3.5 summarizes the results obtained in this setting. As we could expect we found that the three smooth models (Poisson, PRIDE and Smooth-CAR) are equivalent. Poisson P -spline model fits the non-linear trend, which it is equivalent to obtain values of κ^{-1} close to zero in the PRIDE model (i.e. there is no overdispersion), and for the Smooth-CAR model, both structured (σ_s) and unstructured (κ^{-1}) variability are very small. The intrinsic CAR model, estimates the (non-linear) trend by the linear component $(X\hat{\beta})$, and a random effect b , that assumes all the variability is due to the spatial CAR structure (defined by the neighborhood matrix Q^-). Figure 3.11, shows the empirical $\log\text{MSE}$ boxplots for the fitted models. It is worth noticing that CAR model, has

worst values of the logMSE than the other alternatives.

Scenario 2: Smooth spatial non-linear trend with overdispersion

In this scenario, we considered a non-linear trend with overdispersion. This setting consists in adding to the non-linear trend a random noise through a random effect, i.e. $\gamma \sim \mathcal{N}(0, \kappa^{-1} \mathbf{I})$. We considered two situations: low and high overdispersion, with values of $\kappa^{-1} = \{0.0625, 1\}$. As shown in Figure 3.12, Poisson P -spline model fit has the highest values of the MSE, since it fails to capture the presence of overdispersion. Table 3.6 summarizes the estimated parameters for each of the fitted models. The PRIDE model, captures the smooth trend (by $\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\alpha}$) and the overdispersion through $\hat{\kappa}^{-1}$. The Smooth-CAR model estimates close to zero values for the structured variability $\hat{\sigma}_s$, and estimates the overdispersion $\hat{\kappa}^{-1}$ accurately in both situations of low and high overdispersion. Thus, Smooth-CAR model is reduced to PRIDE model.

Finally, the results obtained for the CAR model fit, shows us that in the presence of a non-linear trend, the CAR model attributes all the source of variability to the spatial component, and no variability for the overdispersion. This reflects that the CAR model misspecifies the underlying true source of variability. The results of the MSE values in Figure 3.12 for CAR model are also worst than PRIDE and Smooth-CAR models.

Table 3.6: Mean and standard errors of estimated parameters for Scenario 2, with $\kappa = \{0.0625, 1\}$.

<i>True</i>	Model	$\hat{\sigma}_s$	$\hat{\kappa}^{-1}$
$\kappa^{-1} = 0.0625$	<i>PRIDE</i>	-	0.0642
		-	(0.0261)
	<i>Smooth-CAR*</i>	10^{-4}	0.0623
		(< 10^{-3})	(< 10^{-3})
	<i>CAR*</i>	0.6044	10^{-4}
		(< 10^{-3})	(< 10^{-3})
<i>True</i>	Model	$\hat{\sigma}_s$	$\hat{\kappa}^{-1}$
$\kappa^{-1} = 1$	<i>PRIDE</i>	-	1.0120
		-	(0.0125)
	<i>Smooth-CAR*</i>	10^{-4}	1.0122
		(< 10^{-3})	(0.0112)
	<i>CAR*</i>	0.8635	10^{-4}
		(< 10^{-3})	(< 10^{-3})

* Models fitted with Besag's model CAR covariance structure.

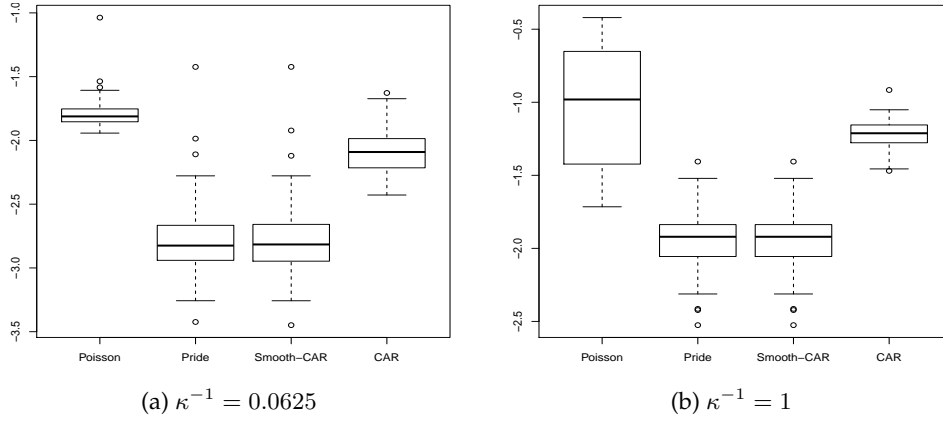


Figure 3.12: $\log(\text{MSE})$ comparison of Poisson, PRIDE, Smooth-CAR and CAR models in scenario 2 with $R = 100$.

Scenario 3: Smooth spatial non-linear trend with CAR random effect

In this scenario, we considered the simulation of a spatial non-linear trend with a structured variability given by a CAR random effect. We considered a spatially structured variability with $\sigma_s = \{0.25, 1\}$. We simulated different situations: (i) low spatial dependence ($\phi = 0.25$), (ii) a situation where spatial variability and overdispersion have the same weight ($\phi = 0.5$), and (iii) the situation where the spatial component has a greater weight ($\phi = 0.75$).

In this setting, we have considered two sources of spatial variability: a non-linear spatial trend, and a CAR structured spatial variability. The PRIDE model does not differentiate between the two sources of variability, and tends to capture the spatial component by means of the bivariate P -spline and attributes the unstructured variability through $\hat{\kappa}^{-1}$, that is equivalent to consider a true value of $\kappa^{-1} \approx \sigma_s^2(1 - \phi)$ (see Table 3.7). The Smooth-CAR model, tends to capture the non-linear trend by the spline and the CAR structure through the random effect for the combinations of parameters simulated. However, in real situations, both structures (non-linear trend and structured variability) might not be completely identifiable, since both large and small-scale variability are not distinct. Finally, in the presence of non-linear trends, as in scenarios 1 and 2, CAR model estimates higher values of the weight parameter ϕ , since tends to attribute the non-linear trend to the structured variability. Figure 3.15 shows the boxplots of the $\log\text{MSE}$ values, we have not included the performance of the Poisson P -spline model, since it gives very large values for the MSE, that may distort the comparison with the other alternative models.

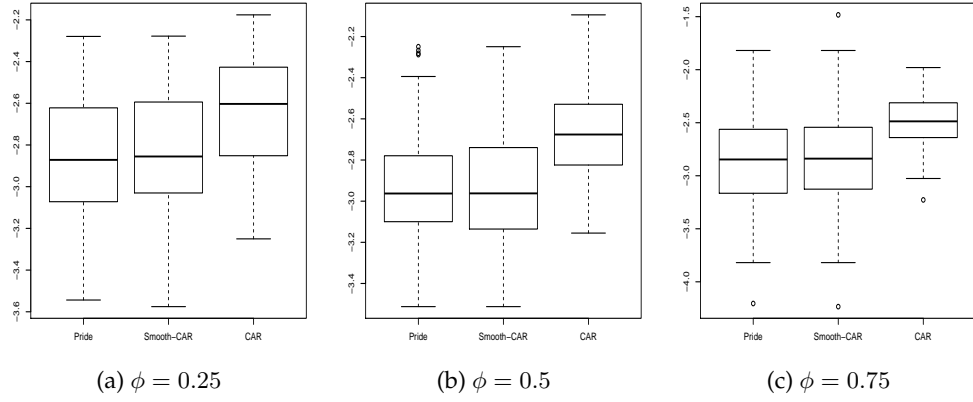


Figure 3.13: $\log(\text{MSE})$ comparison of PRIDE, Smooth-CAR and CAR models in scenario 3 with $R = 100$ and $\sigma_s = 0.25$.

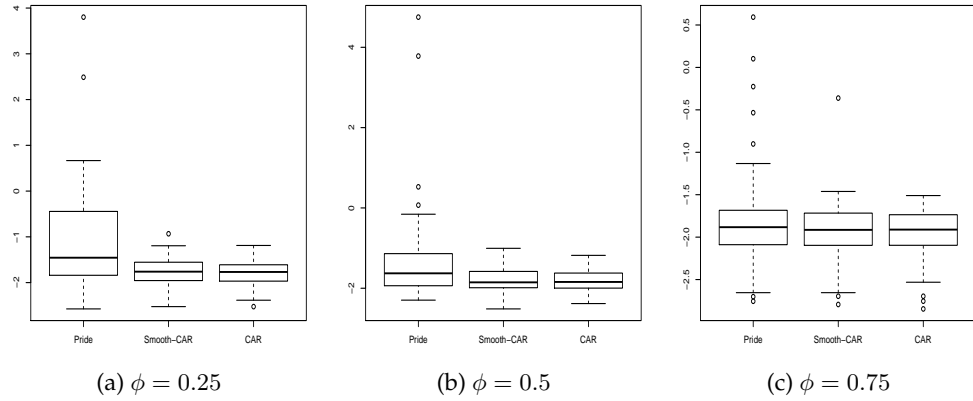


Figure 3.14: $\log(\text{MSE})$ comparison of PRIDE, Smooth-CAR and CAR models in scenario 3 with $R = 100$ and $\sigma_s = 1$.

Scenario 4: Smooth spatial linear trend with CAR random variability

In this scenario, the true linear predictor consists in a linear trend with a CAR random effect (with Dean's model covariance). We considered for the spatially structured variability with $\sigma_s = \{0.25, 0.75\}$, and as in scenario 3, we simulated different situations considering $\phi = \{0.25, 0.5, 0.75\}$.

In this setting, the PRIDE model tends to model the linear predictor and the spatial variability (both linear trend and structured variability) by the spline component, and attributes the extra-variability to overdispersion by the estimation of $\kappa^{-1} \approx \sigma_s^2(1 - \phi)$. In this case, although the true simulated trend is linear, the fitted trend of the PRIDE P -spline ($\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\alpha}$) is not linear, i.e. the smoothing parameters λ_1 , and λ_2 do not tend to infinity, since the random part ($\mathbf{Z}\hat{\alpha}$), is forced to capture the spatial structured

Table 3.7: Mean and standard errors of estimated parameters for Scenario 3, with $\sigma_s = \{0.25, 1\}$, and $\phi = \{0.25, 0.5, 0.75\}$.

<i>True</i>		Model	$\hat{\sigma}_s$	$\hat{\kappa}^{-1}$	$\hat{\phi}$
$\sigma_s = 0.25$	$\phi = 0.25$	<i>PRIDE</i>	-	0.0453	-
			-	(0.0398)	-
		<i>Smooth-CAR</i>	0.2750	-	0.3461
			(0.1801)	-	(0.4455)
		<i>CAR</i>	0.6722	-	0.9677
			(0.4455)	-	(0.0626)
	$\phi = 0.5$	<i>PRIDE</i>	-	0.0314	-
			-	(0.0339)	-
		<i>Smooth-CAR</i>	0.2565	-	0.4203
			(0.1391)	-	(0.4560)
		<i>CAR</i>	0.6771	-	0.9797
			(0.1059)	-	(0.0785)
	$\phi = 0.75$	<i>PRIDE</i>	-	0.0139	-
			-	(0.0274)	-
		<i>Smooth-CAR</i>	0.2342	-	0.7277
			(0.1815)	-	(0.4859)
		<i>CAR</i>	0.6434	-	0.9877
			(0.0956)	-	(0.0444)
<i>True</i>					
$\sigma_s = 1$	$\phi = 0.25$	<i>PRIDE</i>	-	0.7024	-
			-	(0.3651)	-
		<i>Smooth-CAR</i>	1.0855	-	0.3644
			(0.3272)	-	(0.3851)
		<i>CAR</i>	1.2682	-	0.5749
			(0.3172)	-	(0.3172)
	$\phi = 0.5$	<i>PRIDE</i>	-	0.5172	-
			-	(0.1221)	-
		<i>Smooth-CAR</i>	1.0225	-	0.4770
			(0.2841)	-	(0.3729)
		<i>CAR</i>	0.3887	-	0.7269
			(0.2700)	-	(0.2476)
	$\phi = 0.75$	<i>PRIDE</i>	-	0.2553	-
			-	(0.1246)	-
		<i>Smooth-CAR</i>	0.9758	-	0.7330
			(0.1340)	-	(0.3721)
		<i>CAR</i>	0.3492	-	0.8412
			(0.2674)	-	(0.2236)

correlation ($\sigma_s^2 \phi \mathbf{Q}^-$). The Smooth-CAR model, gives large values of the smoothing parameters of the spline, that tend to infinity (i.e. $\lambda_1, \lambda_2 \rightarrow \infty$), leading to a linear trend, and the structure is captured by the CAR component. Thus, the Smooth-CAR model gives similar results to the CAR model.

Table 3.8: Mean and standard errors of estimated parameters for Scenario 4, with $\sigma_s = \{0.25, 0.75\}$, and $\phi = \{0.25, 0.5, 0.75\}$.

True		Model	$\hat{\sigma}_s$	$\hat{\kappa}^{-1}$	$\hat{\phi}$
$\sigma_s = 0.25$	$\phi = 0.25$	PRIDE	-	0.0497	-
			-	(0.0149)	-
		Smooth-CAR	0.2625 (0.0685)	- -	0.2881 (0.3411)
		CAR	0.2642 (0.0675)	- -	0.2973 (0.3400)
	$\phi = 0.5$	PRIDE	-	0.0283	-
			-	(0.0127)	-
Smooth-CAR		0.2433 (0.0724)	- -	0.5017 (0.3880)	
	CAR	0.2495 (0.0703)	- -	0.5122 (0.3778)	
$\phi = 0.75$	PRIDE	-	0.0358	-	
			(0.0104)	-	
		Smooth-CAR	0.2356 (0.0658)	- -	0.7325 (0.3356)
		CAR	0.2462 (0.0841)	- -	0.7421 (0.3512)
	True				
	$\sigma_s = 0.75$	$\phi = 0.25$	PRIDE	-	0.4544
			-	(0.0936)	-
Smooth-CAR			0.7452 (0.1278)	- -	0.2316 (0.2520)
		CAR	0.7465 (0.1274)	- -	0.2328 (0.2528)
$\phi = 0.5$		PRIDE	-	0.2842	-
			-	(0.0801)	-
	Smooth-CAR	0.7264 (0.1552)	- -	0.5040 (0.3187)	
	CAR	0.7669 (0.1908)	- -	0.5126 (0.3155)	
$\phi = 0.75$	PRIDE	-	0.1436	-	
			-	(0.0636)	-
		Smooth-CAR	0.7026 (0.1825)	- -	0.6611 (0.3210)
		CAR	0.7119 (0.1772)	- -	0.6819 (0.3116)

Conclusions of the simulation study

The simulation study performed in this Section, allowed us to evaluate the performance of PRIDE and Smooth-CAR models in different situations, with respect to the widely used CAR model. The Smooth-CAR model, is by construction an *hybrid-model*, and is

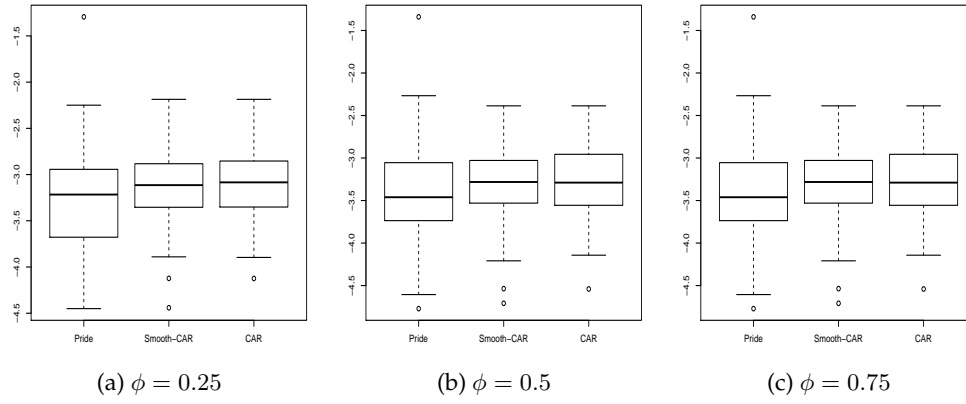


Figure 3.15: $\log(\text{MSE})$ comparison of PRIDE, Smooth-CAR and CAR models in scenario 4 with $R = 100$ and $\sigma_s = 0.25$.

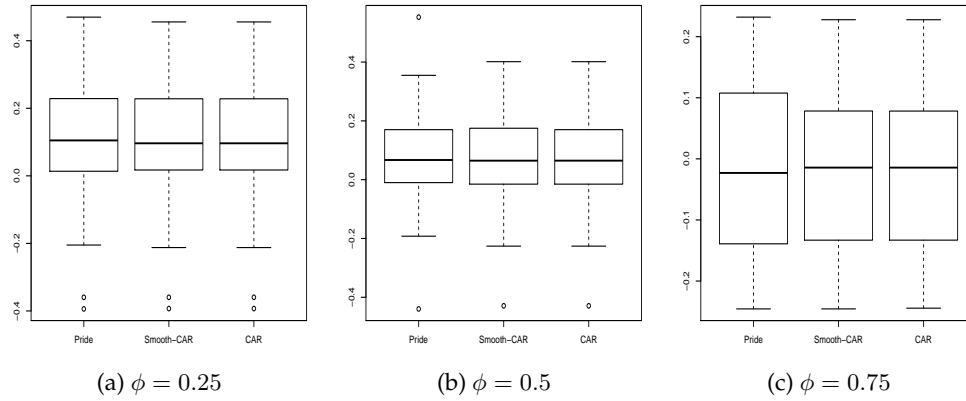


Figure 3.16: $\log(\text{MSE})$ comparison of PRIDE, Smooth-CAR and CAR models in scenario 4 with $R = 100$ and $\sigma_s = 0.75$.

able to capture the different sources of variability in all of the proposed scenarios. The results obtained shows us that the Smooth-CAR has similar performance to the model from which we have simulated the true linear predictor.

When the true linear predictor explicitly incorporates a CAR random effect (scenarios 3 and 4), the Smooth-CAR is capable to capture and separate both sources of variability: (i) the large scale variability, given by the spatial linear/non-linear trends, and (ii) the small-scale variability defined by the CAR component. However, both sources might not be identifiable in real applications where the underlying variability of the spatial stochastic processes may depend on different sources. The PRIDE model is not able to separate between large and small-scale variability, and only distinguish between large-scale (spatial trend estimated by mean of the splines) and unstructured variability or overdispersion. This leads in general to biased estimates and inefficient

analysis. However, despite of this, the PRIDE model might be a good alternative when for example the number of regions is very large and the estimation of a CAR structure covariance (G_b) of dimensions $n \times n$, is computationally intensive, and a feasible solution is to consider low-rank models as spatial P -splines.

The simulation study also led us to interesting questions for further research that have been already studied in the literature of disease mapping applications using CAR models. Several authors have proposed tests for the presence of structured heterogeneity for the hierarchical CAR models (see [Dean et al., 2001](#); [MacNab and Dean, 2000](#); [Ugarte et al., 2005](#)). [Dean et al. \(2001\)](#) proposed the use of score tests for testing for the unstructured heterogeneity. In the context of GLMMs these tests are based on testing the variance components of the model (see [Self and Liang, 1987](#); [Lin, 1997](#)). More formally, using the reparameterization of the CAR model proposed by Dean, i.e. with covariance $G_b = \sigma_s^2(\phi Q^- + (1 - \phi)I)$, the testing for the unstructured variability would be equivalent to testing:

$$H_0 : \phi = 0 \quad \text{versus} \quad H_1 : \phi > 0.$$

The score test statistic for H_0 is then the penalized quasi-likelihood (PQL) estimating function evaluated at $\hat{\sigma}^2$ and $\phi = 0$ (see [Dean et al., 2001](#); [Ugarte et al., 2005](#), for details). [MacNab and Dean \(2000\)](#) proposed for CAR models a parametric bootstrap test for testing H_0 , however some drawbacks are the computing times in obtaining the bootstrap samples and also that the variability of the estimation of the parameter ϕ is large, that it is not taken into account in these tests.

In the context of the smooth models (PRIDE and Smooth-CAR), the idea of testing for the variance components is also of interest. For instance, in the first two scenarios, where we simulated non-linear trends with and without overdispersion, the test would be:

- Test for overdispersion in scenarios 1 and 2 are:

$$\begin{aligned} H_0 : \kappa^{-1} = 0 & \quad \text{versus} \quad H_1 : \kappa^{-1} > 0 & \text{(PRIDE model)} \\ H_0 : \sigma_s^2, \kappa^{-1} = 0 & \quad \text{versus} \quad H_1 : \sigma_s^2, \kappa^{-1} > 0 & \text{(Smooth-CAR model)} \end{aligned}$$

In scenarios 3 and 4, we might be interested in testing for the linear trend, the test would be equivalent to test if the smoothing parameters λ_1 and λ_2 tend to infinity, i.e.:

- Test for linear trend in scenarios 3 and 4:

$$H_0 : \lambda_1^{-1}, \lambda_2^{-1} = 0 \quad \text{versus} \quad H_1 : \lambda_1^{-1}, \lambda_2^{-1} > 0$$

In the context of P -splines as mixed models, Crainiceanu et al. (2005) and Greven et al. (2008) have developed computationally efficient tests for variance components based on the restricted likelihood test statistic (Self and Liang, 1987, 1995; Stram and Lee, 1994). However, in some situations described above the problem of testing several variance components increases the complexity of the methodology. The implementation of these formal tests for the variance components is a topic of current research in more complex situations. We will discuss the drawbacks and issues in more details for the multidimensional case in Section 4.3 of Chapter 4.

Finally, in spatial applications, we may have additional information as possible covariates to be considered to incorporate to the modelling (e.g. enviromental, epidemiologic or socioeconomic variables). These covariates may influence on the linear predictor as non-linear effects or interactions. In the context of smoothing models, these effects can be included in the context of additive models (Hastie and Tibshirani, 1990) to construct models fo the form:

$$\boldsymbol{\eta} = \underbrace{f(\text{lon}, \text{lat})}_{\text{spatial}} + \underbrace{f(\boldsymbol{x}_3) + f(\boldsymbol{x}_4) + \dots + f(\boldsymbol{x}_k)}_{\text{non-linear additive effects}}.$$

These effects can be easily included as random effects in our unified mixed model approach. We will consider in next Chapter the incorporation of additive and interactions smooth terms.

*"Inspiration does exist,
but it must find you working".
Pablo Picasso*

Chapter 4

Smooth-ANOVA models

This Chapter introduces a new class of multidimensional models based on the smoothing mixed model methodology developed in [Chapter 2](#). This type of models are based on the decomposition of multidimensional smooth functions as additive terms and interactions. Previous works by [Gu and Wahba \(1993\)](#); [Wahba et al. \(1995\)](#), [Wang \(1998a,b\)](#) and more recently the book of [Gu \(2002\)](#), have proposed *Smoothing Spline Analysis-of-Variance* (SS-ANOVA) decompositions. Their models obtain main fixed effects and interactions which can be interpreted as in classical ANOVA setting. This interpretation of multidimensional smoothing is often useful when the interaction effects are as interesting as main effects. However, their use is constrained to the amount of data since they are based on full rank smoothers.

This Chapter is organized as follows, we start with the P -spline representation of an additive model, and show the mixed model formulation using the reparameterization we have introduced in [Chapter 2](#). This additive formulation allows us to extend the model by incorporating interaction terms as a P -spline Smooth-ANOVA models. The construction of this model, and the identifiability problems are discussed in [Section 4.2](#). In real applications, sometimes it is of interest to consider Smooth-ANOVA models that include some terms and ignore others. We called these models *reduced* S-ANOVA. We apply this methodology to the spatio-temporal context in [Section 4.4](#) to study air pollution ozone levels in Europe. We also propose a new computationally efficient methodology in [Section 4.5.1](#), using reduced rank bases for the space-time interaction.

4.1 P -spline additive models

In the context of non-parametric regression, additive models (Stone, 1985; Buja et al., 1989; Friedman and Silverman, 1989) are a useful technique in data analysis. These models represent a response variable y as the sum of k smooth functions of covariates x_1, \dots, x_k . Then an additive model for the response y is defined as:

$$y = f_1(x_1) + f_2(x_2) + \dots + f_k(x_k) + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad (4.1)$$

where f_i are smooth functions of the covariates x_i , for $i = 1, 2, \dots, k$. These models have the attractive feature of modelling the effects of the covariates over the response as a sum of individual effects. However, the assumption of additivity results in some cases very restrictive. Another issue in additive models is the problem of identifiability, due to the fact that model (4.1) contains k smooth functions. The simplest way to avoid this problem is to incorporate an intercept term γ , such that $\mathbb{E}[y] = \gamma$, since, otherwise, the fitted curves \hat{f}_i would be unique only up to a constant. It also implicitly incorporates a *sum to zero constraint* as $\sum_{i=1}^n f_i = 0$, for each k smooth term, in order to make the definitions of the functions unique. The book by Hastie and Tibshirani (1990) present an extensive review of additive models for generalized responses as *generalized additive models* or GAMs. The estimation is done by the so-called *backfitting algorithm*. This method is computationally very efficient but, as an iterative procedure, it does not give explicit expressions for the estimated smooth curves.

The representation of additive models, in the context of the P -splines, was introduced by Marx and Eilers (1998) (P -GAMs). They proposed the use of P -splines as a low-rank smoothers, and all smooth terms are estimated simultaneously using a modified version of the scoring algorithm; and the backfitting procedure is avoided. The additive model (4.1) can be written as:

$$y = \gamma \mathbf{1} + B_1 \theta_1 + \dots + B_k \theta_k + \epsilon = B \theta + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2), \quad (4.2)$$

where we incorporate the intercept term γ , with B -spline regression basis:

$$B = (\mathbf{1}_n : B_1 : \dots : B_k), \quad (4.3)$$

with vector of regression coefficients $\theta' = (\gamma, \theta_1, \dots, \theta_k)'$. The penalty matrix P over the regression coefficients, except the constant γ , is blockdiagonal of the form:

$$P = \text{blockdiag}(0, P_1, \dots, P_k), \text{ with } P_i = \lambda_i D_i' D_i \quad (4.4)$$

Given values of the smoothing parameters λ_i (for $i = 1, \dots, k$), we obtain the vector of regression coefficients $\boldsymbol{\theta}$ by minimising the penalized sum of squares in (2.1) and obtain the system of equations:

$$\begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{0}' & \dots & \mathbf{0}' \\ \mathbf{0} & \mathbf{B}_1\mathbf{B}_1 + \lambda_1\mathbf{D}_1'\mathbf{D}_1 & \dots & \mathbf{B}_1'\mathbf{B}_k \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{B}_k'\mathbf{B}_1 & \dots & \mathbf{B}_k\mathbf{B}_k + \lambda_k\mathbf{D}_k'\mathbf{D}_k \end{bmatrix} \begin{bmatrix} \gamma \\ \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_k \end{bmatrix} = \begin{bmatrix} \mathbf{1}' \\ \mathbf{B}_1' \\ \vdots \\ \mathbf{B}_k' \end{bmatrix} \mathbf{y}. \quad (4.5)$$

This solution (4.5) is equivalent to the set of normal equations as in [Hastie and Tibshirani \(1987\)](#). However, due to the fact that the columns of the \mathbf{B}_k sum to one, the B -spline basis in (4.3) is not full column rank, in fact $\text{rank}(\mathbf{B}) = 1 - k + \sum_{i=1}^k c_i$, where c_i is the number of columns of the i^{th} , and $\mathbf{B}'\mathbf{B}$ is singular. [Marx and Eilers \(1998\)](#) proposed the use of a small ridge penalty in the system of equations in (4.5) to solve this problem. An alternative is the use of a generalized inverse to avoid the singularity.

Several authors have extended the GAMs formulation as mixed models (GAMMs), for example [Ruppert et al. \(2003\)](#), [Aerts et al. \(2002\)](#), or [Coull et al. \(2001a\)](#) use truncated polynomials as regression basis. [Durbán and Currie \(2003\)](#) present a mixed model formulation using B -splines as basis functions in the presence of correlated errors and estimate the smoothing and correlation parameters by REML. They avoid the identifiability problem by centering the B -spline regression basis (this might be a problem if the size of the basis is large). In next Section, we follow a similar approach, using the reparameterization proposed in [Chapter 2](#) and imposing constraints on the P -spline coefficients.

4.1.1 Smoothing additive mixed models

For illustrative purposes, let us consider the case of two regressors ($k = 2$). The P -spline representation as an additive model (4.1) is given by:

$$\mathbb{E}[\mathbf{y}] = \gamma + f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) = \mathbf{B}\boldsymbol{\theta}, \quad (4.6)$$

where the B -spline regression basis \mathbf{B} is defined as:

$$\mathbf{B} = (\mathbf{1}_n : \mathbf{B}_1 : \mathbf{B}_2), \quad (4.7)$$

of dimension $n \times (1 + c_1 + c_2)$, where $\mathbf{1}_n$ is a column vector of ones of length $n \times 1$, and vector of regression coefficients is $\boldsymbol{\theta} = (\gamma, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$. We impose a penalty on the coefficients $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and leave the constant γ unpenalized. We use a block-diagonal

penalty matrix:

$$\mathbf{P} = \text{blockdiag}(0, \lambda_1 \mathbf{D}'_1 \mathbf{D}_1, \lambda_2 \mathbf{D}'_2 \mathbf{D}_2). \quad (4.8)$$

The regression matrix (4.7) is not of full column rank ($\text{rank}(\mathbf{B}) = c_1 + c_2$). Now, we will avoid this problem by transforming the original B -splines basis into a non-singular new basis. We follow the same procedure as in Section 2.2.1, and apply the SVD of (4.8) to find an appropriate transformation and formulate the model (4.6) into a mixed model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}$, with $\boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{G})$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$.

Proposition 4.1. *The transformation matrix \mathbf{T} for the additive model case with $k = 2$, and a second order penalty (i.e. $q_1 = q_2 = 2$), is the partitioned matrix defined as:*

$$\mathbf{T} = [\mathbf{T}_n : \mathbf{T}_s] = \left[\begin{array}{ccc|cc} 1 & \cdots & 0 & 0 & \\ \vdots & \mathbf{u}_{1n}^{(2)} & \vdots & \mathbf{U}_{1s} & \\ 0 & & \mathbf{u}_{2n}^{(2)} & & \mathbf{U}_{2s} \end{array} \right], \quad (4.9)$$

of dimension $(1 + c_1 + c_2) \times (c_1 + c_2 - 1)$, where $\mathbf{u}_{1n}^{(2)}$ and $\mathbf{u}_{2n}^{(2)}$ are the second columns of \mathbf{U}_{1n} and \mathbf{U}_{2n} , and \mathbf{U}_{1s} and \mathbf{U}_{2s} are the eigenvectors corresponding to the positive eigenvalues of the SVD over the penalty matrix in (4.8).

Proof. Given the additive model (4.6) with basis (4.7) and penalty (4.8), we can define the partitioned matrix $\check{\mathbf{T}}$, constructed block-diagonal submatrices given by:

$$\check{\mathbf{T}} = [\check{\mathbf{T}}_n : \check{\mathbf{T}}_s] = \left[\begin{array}{ccc|cc} 1 & \cdots & 0 & 0 & \cdots \\ \vdots & \mathbf{U}_{1n} & & \mathbf{U}_{1s} & \\ 0 & & \mathbf{U}_{2n} & & \mathbf{U}_{2s} \end{array} \right], \quad (4.10)$$

where the first entry corresponds to the constant term. The submatrices \mathbf{U}_{kn} and \mathbf{U}_{ks} (for $k = 1, 2$) are the eigenvectors corresponding to the zero and non-zero eigenvalues of the SVD over the penalty matrix (4.8). The dimension of the transformation matrix (4.10) is $(1 + c_1 + c_2) \times (1 + c_1 + c_2)$. Then, if we reparameterize the model basis as $\mathbf{B}\check{\mathbf{T}} = [\mathbf{X} : \mathbf{Z}]$, we obtain the fixed part as:

$$\begin{aligned} \mathbf{X} = \mathbf{B}\check{\mathbf{T}}_n &= [\mathbf{1}_n : \mathbf{B}_1 : \mathbf{B}_2] \left[\begin{array}{ccc} 1 & \cdots & 0 \\ \vdots & \mathbf{U}_{1n} & \vdots \\ 0 & & \mathbf{U}_{2n} \end{array} \right] = \\ &= [\mathbf{1}_n : \mathbf{B}_1 \mathbf{U}_{1n} : \mathbf{B}_2 \mathbf{U}_{2n}] = \\ &= [\mathbf{1}_n : \mathbf{X}_1 : \mathbf{X}_2]. \end{aligned} \quad (4.11)$$

For a second order penalty, we can take $\mathbf{X}_1 = [\mathbf{1}_n : \mathbf{x}_1]$ and $\mathbf{X}_2 = [\mathbf{1}_n : \mathbf{x}_2]$. Given these definitions, it can be seen in (4.11) that the column vector of ones $\mathbf{1}_n$ appears more than once. This is the cause of the linear dependency among the columns of \mathbf{X} . We avoid the linear dependency in (4.11) simply removing the column vectors of ones in \mathbf{X}_1 and \mathbf{X}_2 . Thus, we replace the fixed effects matrix (4.11) by:

$$\mathbf{X} = [\mathbf{1}_n : \mathbf{x}_1 : \mathbf{x}_2]. \quad (4.12)$$

The random part is obtained as:

$$\begin{aligned} \mathbf{Z} = \mathbf{B}\check{\mathbf{T}}_s &= [\mathbf{1}_n : \mathbf{B}_1 : \mathbf{B}_2] \begin{bmatrix} 0 & \cdots \\ \mathbf{U}_{1s} & \\ & \mathbf{U}_{2s} \end{bmatrix} = [\mathbf{B}_1\mathbf{U}_{1s} : \mathbf{B}_2\mathbf{U}_{2s}] = \\ &= [\mathbf{Z}_1 : \mathbf{Z}_2]. \end{aligned} \quad (4.13)$$

Note that, the procedure of removing the first column vectors in the matrices \mathbf{X}_1 and \mathbf{X}_2 , in terms of the transformation of the bases implies removing the equivalent column vectors of the submatrices \mathbf{U}_{1n} and \mathbf{U}_{2n} . Hence, the transformation matrix used to avoid the identifiability problem of the additive model, consists in removing the first eigenvectors of \mathbf{U}_{1n} and \mathbf{U}_{2n} in $\check{\mathbf{T}}$ defined in (4.10). This is the matrix defined in (4.9). ■

Theorem 4.1. *The penalty for an additive mixed model with two regressors and second order penalty ($q_1 = q_2 = 2$) is the block-diagonal matrix:*

$$\mathbf{F} = \text{blockdiag}(\lambda_1 \tilde{\Sigma}_1, \lambda_2 \tilde{\Sigma}_2) \text{ of size } (c_1 + c_2 - 4) \times (c_1 + c_2 - 4), \quad (4.14)$$

where $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$ are the positive eigenvalues of the penalty matrices $\mathbf{D}'_1\mathbf{D}_1$ and $\mathbf{D}'_2\mathbf{D}_2$.

Proof. For given matrix \mathbf{T}_s defined in (4.9), and penalty matrix \mathbf{P} for the additive model in (4.8) with ($q_1 = q_2 = 2$), the block-diagonal matrix is obtained by $\mathbf{F} = \mathbf{T}'_s\mathbf{P}\mathbf{T}_s$ as in 2.1, i.e.:

$$\begin{aligned} \mathbf{F} = \mathbf{T}'_s\mathbf{P}\mathbf{T}_s &= \begin{pmatrix} 0 & \mathbf{U}'_{1s} & \\ & & \mathbf{U}'_{2s} \end{pmatrix} \begin{pmatrix} 0 & & \\ \lambda_1 \mathbf{D}'_1\mathbf{D}_1 & & \\ & \lambda_2 \mathbf{D}'_2\mathbf{D}_2 & \end{pmatrix} \begin{pmatrix} 0 & \\ \mathbf{U}_{1s} & \\ & \mathbf{U}_{2s} \end{pmatrix} = \\ &= \begin{pmatrix} \lambda_1 \mathbf{U}'_{1s}\mathbf{D}'_1\mathbf{D}_1\mathbf{U}_{1s} & & \\ & \lambda_2 \mathbf{U}'_{2s}\mathbf{D}'_2\mathbf{D}_2\mathbf{U}_{2s} & \end{pmatrix}, \end{aligned}$$

where $\tilde{\Sigma}_k = \mathbf{U}'_{ks}\mathbf{D}'_k\mathbf{D}_k\mathbf{U}_{ks}$ for $k = 1, 2$. ■

Remark 4.1. We can generalize the result in [Theorem 4.1](#) for an additive model with $k > 2$, as

$$\mathbf{F} = \bigoplus_{i=1}^k \lambda_i \tilde{\Sigma}_i = \text{blockdiag}(\lambda_1 \tilde{\Sigma}_1, \dots, \lambda_k \tilde{\Sigma}_k).$$

Given the transformation matrix \mathbf{T} for the additive model, defined in (4.9), we are able to transform the original B -spline basis \mathbf{B} into a non-singular basis of full column rank (equal to $c_1 + c_2$), by $\mathbf{BT} = [\mathbf{X} : \mathbf{Z}]$, and therefore obtain the blockdiagonal penalty matrix \mathbf{F} . The estimation of the variance components can be done using REML and the fixed and random effects coefficients are obtained using the standard mixed model equations. The extension to non-Gaussian responses as generalized additive models (GAMs), is then straightforward (as we already shown in [Section 2.2.2](#)). Using this reparameterization of the additive model, the hat-matrices and the effective dimension for each smooth term are easily obtained. For example, for given values of σ and λ_i , and $i = 1, \dots, k$. We have the hat-matrix for $f_i(\mathbf{x}_i)$, can be directly computed as:

$$\mathbf{H}_i = \mathbf{C}_i \begin{bmatrix} \mathbf{x}_i' \mathbf{x}_i & \mathbf{x}_i' \mathbf{Z}_i \\ \mathbf{Z}_i' \mathbf{x}_i & \mathbf{Z}_i' \mathbf{Z}_i + \mathbf{G}_i^{-1} \end{bmatrix}^{-1} \mathbf{C}_i', \quad (4.15)$$

where $\mathbf{C}_i = [\mathbf{x}_i : \mathbf{Z}_i]$, and $\mathbf{G}_i = \sigma^2 \mathbf{F}_i^{-1}$, and $\mathbf{F}_i = \lambda_i \tilde{\Sigma}_i$. The effective dimension (ED) associated to each smooth term $f_i(\mathbf{x}_i)$ is:

$$\text{ED}_i = \text{trace}(\mathbf{H}_i) = \text{trace} \left\{ \begin{bmatrix} \mathbf{x}_i' \mathbf{x}_i & \mathbf{x}_i' \mathbf{Z}_i \\ \mathbf{Z}_i' \mathbf{x}_i & \mathbf{Z}_i' \mathbf{Z}_i + \mathbf{G}_i^{-1} \end{bmatrix}^{-1} \mathbf{C}_i' \mathbf{C}_i \right\}. \quad (4.16)$$

Then, we have that $\hat{f}(\mathbf{x}_i) = \mathbf{H}_i \mathbf{y}$, and:

$$\mathbb{E}[\mathbf{y}] = \gamma + \sum_{i=1}^k \hat{f}_i(\mathbf{x}_i) = \gamma + \sum_i^k \mathbf{H}_i \mathbf{y}.$$

Thus, the total effective dimension of an additive model is $\text{ED} = 1 + \sum_{i=1}^k \text{ED}_i$, where 1 degree corresponds to the constant term.

GAMs are a very useful modelling tool when the main effects of the covariates are simply added together to obtain a joint effect on the response. This assumption, where the interactions between the covariates are completely ignored is too restrictive in many situations. In next Section, we extend the smoothing additive mixed models methodology to the incorporation of interactions. We use the term *Smooth-ANOVA models*.

4.2 P-spline smooth-ANOVA models

Sometimes the interest lies in fitting complex multidimensional models with functional form given by

$$\mathbb{E}[\mathbf{y}] = \gamma + \sum_{i=1}^k f_i(\mathbf{x}_i) + \sum_{i<j} f_{ij}(\mathbf{x}_i, \mathbf{x}_j) + \cdots + f_{1,\dots,k}(\mathbf{x}_1, \dots, \mathbf{x}_k), \quad (4.17)$$

where γ is a constant term, f_i are additive univariate functions of the i^{th} covariate, f_{ij} a two-dimensional interaction smooth function of the pair of covariates $(\mathbf{x}_i, \mathbf{x}_j)$, and so on until a k^{th} order interaction. As addressed by [Chen \(1993\)](#), these type of models can be seen as a functional version of *Analysis-Of-Variance* (ANOVA). Using this terminology, model (4.17) is the sum of smooth functions of *main effects* and *two-way interactions*, *three-way interactions*, and so on. However, higher-order interactions are less interpretable and more difficult to estimate due to the curse of dimensionality. In general, higher-order interactions are usually ignored in order to reduce the model complexity. As in a classical ANOVA model, in most situations only main effects and second order interactions are considered in practice. Note that, additive models presented in [Section 4.1](#) are a special case of model (4.17) when only main effects are included (more details in the Smoothing Splines ANOVA (SS-ANOVA) can be found in the book by [Gu \(2002\)](#)).

In this Section, we present a new approach based on P -splines for the ANOVA-type decomposition of multidimensional smooth functions. The identifiability problems are avoided using the mixed model reparameterization shown in the previous Chapter. To illustrate our procedure, we will extend the additive model representation of a P -spline presented in [Section 4.1.1](#) to models which include the interaction term based on Tensor product of individual bases.

4.2.1 Smoothing additive mixed models with interactions

Let us consider the two-dimensional case when data are in an array structure. The data vector \mathbf{y} of length $n \times 1$, where $n = n_1 n_2$, and the regressors: $\mathbf{x}_1 = (x_{1i}, \dots, x_{1n_1})'$ and $\mathbf{x}_2 = (x_{2j}, \dots, x_{2n_2})'$, for $i = 1, \dots, n_1$ $j = 1, \dots, n_2$. The Smooth-ANOVA model is given by:

$$\mathbb{E}[\mathbf{y}] = \gamma + f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + f_{1,2}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{B}\boldsymbol{\theta}, \quad (4.18)$$

where the B -spline regression basis \mathbf{B} is defined as:

$$\mathbf{B} = (\mathbf{1}_n : \mathbf{1}_{n_2} \otimes \mathbf{B}_1 : \mathbf{B}_2 \otimes \mathbf{1}_{n_1} : \mathbf{B}_2 \otimes \mathbf{B}_1), \quad (4.19)$$

of dimension $n \times (1 + c_1 + c_2 + c_1 c_2)$, and where $\mathbf{1}_{n_1}$ and $\mathbf{1}_{n_2}$ are column vectors of ones of length n_1 and n_2 respectively, and the vector of regression coefficients $\boldsymbol{\theta}' = (\gamma, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_s)'$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the $c_1 \times 1$ and $c_2 \times 1$, vector of coefficients for the main effects and $\boldsymbol{\theta}_s$, of length $c_1 c_2 \times 1$, for the interaction. Then, model (4.18) can be written as:

$$\mathbb{E}[\mathbf{y}] = \gamma \mathbf{1}_n + (\mathbf{1}_{n_2} \otimes \mathbf{B}_1) \boldsymbol{\theta}_1 + (\mathbf{B}_2 \otimes \mathbf{1}_{n_1}) \boldsymbol{\theta}_2 + (\mathbf{B}_2 \otimes \mathbf{B}_1) \boldsymbol{\theta}_s.$$

In matrix notation, we can arrange the response vector \mathbf{y} in a $n_1 \times n_2$ matrix, \mathbf{Y} and use the array notation as:

$$\mathbb{E}[\mathbf{Y}] = \gamma \mathbf{1}_n + \mathbf{B}_1 \boldsymbol{\Theta}_1 \mathbf{1}_{n_2}' + \mathbf{1}_{n_1} \boldsymbol{\Theta}_2 \mathbf{B}_2' + \mathbf{B}_1 \boldsymbol{\Theta}_s \mathbf{B}_2', \quad (4.20)$$

where $\boldsymbol{\Theta}_1$ is of dimension $c_1 \times 1$, $\boldsymbol{\Theta}_2$ is $1 \times c_2$, and $\boldsymbol{\Theta}_s$ is the $c_1 \times c_2$ matrix of coefficients for the interaction, such that $\text{vec}(\boldsymbol{\Theta}_1) = \boldsymbol{\theta}_1$, $\text{vec}(\boldsymbol{\Theta}_2) = \boldsymbol{\theta}_2$ and $\text{vec}(\boldsymbol{\Theta}_s) = \boldsymbol{\theta}_s$. The penalty has a block-diagonal structure of the form:

$$\mathbf{P} = \begin{pmatrix} 0 & \cdots & & \\ \vdots & \lambda_1 \mathbf{D}_1' \mathbf{D}_1 & & \\ & & \lambda_2 \mathbf{D}_2' \mathbf{D}_2 & \\ & & & \tau_2 \mathbf{D}_2' \mathbf{D}_2 \otimes \mathbf{I}_{c_1} + \tau_1 \mathbf{I}_{c_2} \otimes \mathbf{D}_1' \mathbf{D}_1 \end{pmatrix}, \quad (4.21)$$

of dimension $(1 + c_1 + c_2 + c_1 c_2) \times (1 + c_1 + c_2 + c_1 c_2)$, where each block corresponds to the penalty over each of the coefficients of the model. The penalty matrix (4.21) includes one-dimensional penalties of the additive smooth terms with smoothing parameters λ_1 and λ_2 , and a two-dimensional penalty for the interaction term (2.63) with τ_1 and τ_2 . However, as in the additive model case, the regression matrix (4.19) is not full rank, (in fact $\text{rank}(\mathbf{B}) = c_1 c_2$), so there are $(1 + c_1 + c_2)$ linearly dependent columns. In other words, some elements of the basis $\mathbf{1}_{n_2} \otimes \mathbf{B}_1$ and $\mathbf{B}_2 \otimes \mathbf{1}_{n_1}$ are included in the basis for the interaction $\mathbf{B}_2 \otimes \mathbf{B}_1$. And hence, model (4.18) should be modified in order to preserve the identifiability. The identifiability problem is also reflected in the fact that the penalty matrix in (4.21) is rank deficient. For second order penalty matrices $\mathbf{D}_i' \mathbf{D}_i$, for $i = 1, 2$, the penalty (4.21) has rank $(c_1 + c_2 + c_1 c_2 - 8)$. Wood (2006b) pointed out the need to construct appropriate model bases and penalties, and impose constraints to maintain the model identifiability, and Wood (2006a, chap. 4), suggested the use of the QR decomposition in order to identify numerically any linear dependent columns of model bases and remove them. In contrast, we propose a more elegant way to construct identifiable model bases and penalties, based on the reparameterization shown in Sec-

tion 2.3.2. The mixed model representation of model (4.18) allows us to find that some terms are repeated. As we addressed in Section 4.1.1, we will avoid the problem by removing the column vector of 1's in the fixed effects matrices.

4.2.2 Reparametization of the S-ANOVA model into a mixed model formulation

In order to obtain the reparameterization of model (4.18), we use the SVD of the penalty matrices $D'_k D_k$, and the model matrices X_k , and Z_k , for $k = 1, 2$. The mixed model matrices for the additive terms corresponding to covariates x_1 and x_2 , are:

$$f_1(x_1) \equiv \mathbf{1}_{n_2} \otimes [X_1 : Z_1] = [\mathbf{1}_n : \mathbf{1}_{n_2} \otimes x_1 : \mathbf{1}_{n_2} \otimes Z_1], \text{ and} \quad (4.22)$$

$$f_2(x_2) \equiv [X_2 : Z_2] \otimes \mathbf{1}_{n_1} = [\mathbf{1}_n : x_2 \otimes \mathbf{1}_{n_1} : Z_2 \otimes \mathbf{1}_{n_1}]. \quad (4.23)$$

For the interaction term, we have the mixed model matrices of a two-dimensional model as we showed in (2.68) and (2.69), i.e.:

$$\begin{aligned} f_{1,2}(x_1, x_2) &\equiv X_2 \otimes X_1 : Z_2 \otimes X_1 : X_2 \otimes Z_1 : Z_2 \otimes Z_1 = \\ &\equiv [\mathbf{1}_{n_2} : x_2] \otimes [\mathbf{1}_{n_1} : x_1] : Z_2 \otimes [\mathbf{1}_{n_1} : x_1] : [\mathbf{1}_{n_2} : x_2] \otimes Z_1 : Z_2 \otimes Z_1. \end{aligned} \quad (4.24)$$

As we already showed for the additive model case, this reparameterization allows us to identify the linearly dependent columns in the bases (it can be seen that the columns of (4.22) and (4.23) are already contained in (4.24)). Therefore, we can solve the identifiability problems simply removing the column vectors $\mathbf{1}_{n_1}$ and $\mathbf{1}_{n_2}$ in (4.24), as well as from the additive terms (4.22) and (4.23). Then, the fixed and random effects matrices for model in (4.18) are:

$$X = [\mathbf{1}_n : \mathbf{1}_{n_2} \otimes x_1 : x_2 \otimes \mathbf{1}_{n_1} : x_2 \otimes x_1], \text{ and} \quad (4.25)$$

$$Z = [\mathbf{1}_{n_2} \otimes Z_1 : Z_2 \otimes \mathbf{1}_{n_1} : x_2 \otimes Z_1 : Z_2 \otimes x_1 : Z_2 \otimes Z_1], \quad (4.26)$$

where $[X : Z]$ is of full column rank, $c_1 c_2$. Note that, matrices for model $f_1(x_1) + f_2(x_2) + f_{1,2}(x_1, x_2)$ given in (4.25) and (4.26), are exactly equivalent to those obtained in the two-dimensional case $f_{1,2}(x_1, x_2)$. This is due to the fact that $f_1(x_1) + f_2(x_2)$ can be seen as a particular case of a general function of the two covariates $f_{1,2}(x_1, x_2)$. As we discussed in Section 2.3.2, we can decompose the fitted values of the two-dimensional surface $f_{1,2}(x_1, x_2)$ as a S-ANOVA model. In the S-ANOVA case, we allow for more flexibility, since we explicitly consider an additive model with interactions (with diffe-

rent smoothing parameters λ_1 and λ_2 , τ_1 and τ_2).

We have shown that, the mixed model matrices (4.25) and (4.26) were easily obtained by simple elimination of the repeated columns in the matrices. However, the mixed model representation will not be complete unless we give an expression for the variance-covariance matrix of the mixed model random effects: $\mathbf{G} = \sigma^2 \mathbf{F}^{-1}$. We need to define a transformation \mathbf{T} that takes into account the reduction in the dimension of matrices \mathbf{X} and \mathbf{Z} . We detail the construction of the transformation matrix \mathbf{T} in next Section, and then construct the mixed model penalty \mathbf{F} . We will also demonstrate that this result is equivalent to impose linear constraints over the P -spline regression coefficients.

4.2.3 Transformation matrix in S-ANOVA models

We showed that in the additive mixed model case, the transformation matrix in (4.9) had a block-diagonal structure. Now, for S-ANOVA model in (4.18), the transformation matrix will have also a block-diagonal structure.

Proposition 4.2. *Given the S-ANOVA model (4.18) with model basis (4.19) and second order penalty ($q_1 = q_2 = 2$). The transformation matrix \mathbf{T} , such that we reparameterize the model basis as $\mathbf{BT} = [\mathbf{X} : \mathbf{Z}]$, where $[\mathbf{X} : \mathbf{Z}]$ is of full rank, is the partitioned matrix \mathbf{T} , defined as $\mathbf{T} = [\mathbf{T}_n : \mathbf{T}_s]$, where each sub-matrix is given by:*

$$\mathbf{T}_n = \begin{bmatrix} 1 & \cdots & & 0 \\ \vdots & \mathbf{u}_{1n}^{(2)} & & \\ & & \mathbf{u}_{2n}^{(2)} & \\ 0 & & & \mathbf{u}_{2n}^{(2)} \otimes \mathbf{u}_{1n}^{(2)} \end{bmatrix} \quad \text{and} \quad (4.27)$$

$$\mathbf{T}_s = \begin{bmatrix} 0 & \cdots & & \\ \mathbf{U}_{1s} & & & \\ \vdots & \mathbf{U}_{2s} & & \\ & & \mathbf{u}_{2n}^{(2)} \otimes \mathbf{U}_{1s} : \mathbf{U}_{2s} \otimes \mathbf{u}_{1n}^{(2)} : \mathbf{U}_{2s} \otimes \mathbf{U}_{1s} \end{bmatrix}, \quad (4.28)$$

where \mathbf{T} has dimension $(1 + c_1 + c_2 + c_1 c_2) \times c_1 c_2$, and where $\mathbf{u}_{1n}^{(2)}$ and $\mathbf{u}_{2n}^{(2)}$ are the second columns of \mathbf{U}_{1n} and \mathbf{U}_{2n} , and \mathbf{U}_{1s} and \mathbf{U}_{2s} are the eigenvectors corresponding to the positive eigenvalues of the SVD of $\mathbf{D}'_1 \mathbf{D}_1$ and $\mathbf{D}'_2 \mathbf{D}_2$.

Proof. As shown in 4.1, the procedure of removing the first column of the null part eigenvectors \mathbf{U}_{1n} and \mathbf{U}_{2n} , allows us to remove the linear dependent columns in the mixed model basis. Then, for a second order penalty, given the regression basis \mathbf{B} in (4.19) and the transformation matrix \mathbf{T} in (4.27), we obtain the fixed effects matrix in

(4.25) as $X = BT_n$, i.e.:

$$\begin{aligned} X &= [\mathbf{1}_n : \mathbf{1}_{n_2} \otimes B_1 : B_2 \otimes \mathbf{1}_{n_1} : B_2 \otimes B_1] \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \mathbf{u}_{1n}^{(2)} & \vdots \\ & \mathbf{u}_{2n}^{(2)} & \\ 0 & & \mathbf{u}_{2n}^{(2)} \otimes \mathbf{u}_{1n}^{(2)} \end{bmatrix} = \\ &= [\mathbf{1}_n : \mathbf{1}_{n_2} \otimes B_1 \mathbf{u}_{1n}^{(2)} : B_2 \mathbf{u}_{2n}^{(2)} \otimes \mathbf{1}_{n_1} : B_2 \mathbf{u}_{2n}^{(2)} \otimes B_1 \mathbf{u}_{1n}^{(2)}] = \\ &= [\mathbf{1}_n : \mathbf{1}_{n_2} \otimes \mathbf{x}_1 : \mathbf{x}_2 \otimes \mathbf{1}_{n_1} : \mathbf{x}_2 \otimes \mathbf{x}_1], \end{aligned}$$

where $B_1 \mathbf{u}_{1n}^{(2)}$ and $B_2 \mathbf{u}_{2n}^{(2)}$, were replaced by \mathbf{x}_1 and \mathbf{x}_2 . The random effects matrix in (4.26), is obtained as $Z = BT_s$, i.e.:

$$\begin{aligned} Z &= [\mathbf{1}_n : \mathbf{1}_{n_2} \otimes B_1 : B_2 \otimes \mathbf{1}_{n_1} : B_2 \otimes B_1] \begin{bmatrix} 1 & & & \\ & U_{1s} & & \\ & & U_{2s} & \\ & & & \mathbf{u}_{2n}^{(2)} \otimes U_{1s} : U_{2s} \otimes \mathbf{u}_{1n}^{(2)} : U_{2s} \otimes U_{1s} \end{bmatrix} = \\ &= [\mathbf{1}_n : \mathbf{1}_{n_2} \otimes B_1 U_{1s} : B_2 U_{2s} \otimes \mathbf{1}_{n_1} : B_2 \mathbf{u}_{2n}^{(2)} \otimes B_1 U_{1s} : B_2 U_{2s} \otimes B_1 \mathbf{u}_{1n}^{(2)} : B_2 U_{2s} \otimes B_1 U_{1s}] = \\ &= [\mathbf{1}_n : \mathbf{1}_{n_2} \otimes Z_1 : Z_2 \otimes \mathbf{1}_{n_1} : \mathbf{x}_2 \otimes Z_1 : Z_2 \otimes \mathbf{x}_1 : Z_2 \otimes Z_1], \end{aligned}$$

where $B_1 U_{1s}$ and $B_2 U_{2s}$ is replaced by Z_1 and Z_2 . ■

Remark 4.2. Note that, matrix in the partitioned matrices (4.27) and (4.28) we have accounted for the reduction of the column rank, and in this case T is not orthogonal.

Given the transformation matrix in 4.2, by Theorem 2.1, we can obtain the mixed model penalty F for the S-ANOVA in (4.18).

Theorem 4.2 (Mixed model penalty for the two-dimensional S-ANOVA model). *The mixed model penalty for S-ANOVA model in (4.18) is the block-diagonal defined by:*

$$F = \text{blockdiag}(F_{(1)}, F_{(2)}, F_{(1,2)}), \quad (4.29)$$

where for a second order penalty, we have that (4.29) has size $(c_1 c_2 - 4) \times (c_1 c_2 - 4)$, and where:

$$\begin{aligned} F_{(1)} &= \lambda_1 \tilde{\Sigma}_1, \\ F_{(2)} &= \lambda_2 \tilde{\Sigma}_2, \text{ and} \\ F_{(1,2)} &= \text{blockdiag}(\tau_1 \tilde{\Sigma}_1, \tau_2 \tilde{\Sigma}_2, \tau_1 I_{c_2-2} \otimes \tilde{\Sigma}_1 + \tau_2 \tilde{\Sigma}_2 \otimes I_{c_1-2}). \end{aligned}$$

Proof. Given the matrices T_s and P , defined in (4.28) and (4.21), by (2.40), we obtain F

in (4.29) as:

$$\begin{aligned}
 F = T'_s P T_s &= \begin{bmatrix} 0 & U'_{1s} & \cdots \\ \vdots & U'_{2s} & \\ & \mathbf{u}_{2n}^{(2)'} \otimes U'_{1s} & \\ & U'_{2s} \otimes \mathbf{u}_{1n}^{(2)'} & \\ & U'_{2s} \otimes U'_{1s} & \end{bmatrix} \begin{bmatrix} 0 & \cdots \\ \vdots & \lambda_1 D'_1 D_1 \\ & \lambda_2 D'_2 D_2 \\ & \tau_2 D'_2 D_2 \otimes I_{c_1} + \tau_1 I_{c_2} \otimes D'_1 D_1 \end{bmatrix} \\
 &= \begin{bmatrix} 0 & \cdots \\ U_{1s} & \\ \vdots & U_{2s} \\ & \mathbf{u}_{2n}^{(2)} \otimes U_{1s} : U_{2s} \otimes \mathbf{u}_{1n}^{(2)} : U_{2s} \otimes U_{1s} \end{bmatrix} = \\
 &= \begin{bmatrix} \lambda_1 U'_{1s} D'_1 D_1 U_{1s} \\ \lambda_2 U'_{2s} D'_2 D_2 U_{2s} \\ \tau_2 \mathbf{u}_{2n}^{(2)'} D'_2 D_2 \mathbf{u}_{2n}^{(2)} \otimes U'_{1s} U_{1s} + \tau_1 \mathbf{u}_{2n}^{(2)'} \mathbf{u}_{2n}^{(2)} \otimes U'_{1s} D'_1 D_1 U_{1s} \\ \tau_2 U'_{2s} D'_2 D_2 U_{2s} \otimes \mathbf{u}_{1n}^{(2)'} \mathbf{u}_{1n}^{(2)} + \tau_1 U'_{2s} U_{2s} \otimes \mathbf{u}_{1n}^{(2)} D'_1 D_1 \mathbf{u}_{1n}^{(2)} \\ \tau_1 U'_{2s} U_{2s} \otimes U'_{1s} D'_1 D_1 U_{1s} + \tau_2 U'_{2s} D'_2 D_2 U_{2s} \otimes U'_{1s} U_{1s} \end{bmatrix},
 \end{aligned}$$

using $\tilde{\Sigma}_k = U'_{ks} D'_k D_k U_{ks}$, and $\mathbf{u}_{kn}^{(2)'} D'_k D_k \mathbf{u}_{kn}^{(2)} = 0$, $\mathbf{u}_{kn}^{(2)'} \mathbf{u}_{kn}^{(2)} = 1$ and $U'_{ks} U_{ks} = I_{c_k - q_k}$, for $k = 1, 2$. We obtain the mixed model penalty F in (4.29). ■

Once we have obtain the expression for the block-diagonal mixed model penalty, now our aim is to show that the effect of removing the column of 1's in the fixed effects matrices is equivalent to impose the usual constraints on the model coefficients, i.e., solving the identifiability problems in the mixed model, results in transforming the original coefficients and penalty. This can be proved by recovering the penalty of the original parametrization. However, in the S-ANOVA case, we cannot proceed as in the previous cases, since the transformation matrix T is not orthogonal, and the result in (2.41), i.e. $P = T\Phi T'$, does not satisfies.

Definition 4.1 (Recovered penalty matrix in the S-ANOVA model). We define the “recovered penalty matrix” in the non-transformed S-ANOVA model in (4.18), as the matrix defined by:

$$\check{P} = T\Phi T' = T \underbrace{T' P T}_{\Phi} T' = K P K, \quad (4.30)$$

where $K = T T'$.

Remark 4.3. Note that, in 4.1, we use the symbol \smile , given that T is not orthogonal an then $K \neq I$. Hence, the recovered penalty matrix will be in this case different to P .

Remark 4.4. For a second order penalty, the recovered penalty matrix for S-ANOVA model in (4.18) has $\text{rank}(\check{P}) = c_1 c_2 - 4$.

Proposition 4.3. For the S-ANOVA model in (4.18), and recovered penalty: $\check{P} = \mathbf{K} \mathbf{P} \mathbf{K}$, the matrix \mathbf{K} is a “constrast matrix” that centers the regression coefficients $\boldsymbol{\theta}$, defined as:

$$\mathbf{K} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \mathbf{K}_1 & \\ & & \mathbf{K}_2 \\ 0 & & & \mathbf{K}_2 \otimes \mathbf{K}_1 \end{pmatrix}, \quad (4.31)$$

where \mathbf{K}_1 and \mathbf{K}_2 are centering matrices of order c_1 and c_2 respectively. That is, \mathbf{K}_q is a centering matrix is defined as the square matrix of dimension $c_q \times c_q$, given by:

$$\mathbf{K}_q = \mathbf{I}_{c_q} - \mathbf{1}\mathbf{1}'/c_q, \quad (4.32)$$

and \mathbf{K}_q is symmetric and idempotent.

Proof. Given the matrix \mathbf{T} defined by the sub-matrices (4.27) and (4.28), we have that $\mathbf{K} = \mathbf{T} \mathbf{T}'$, i.e.:

$$\begin{aligned} \mathbf{T} \mathbf{T}' &= [\mathbf{T}_n : \mathbf{T}_s] \begin{bmatrix} \mathbf{T}'_n \\ \mathbf{T}'_s \end{bmatrix} = \mathbf{T}_n \mathbf{T}'_n + \mathbf{T}_s \mathbf{T}'_s = \\ &= \begin{bmatrix} 1 & & & \\ & \mathbf{u}_{1n}^{(2)} \mathbf{u}_{1n}^{(2)'} & & \\ & & \mathbf{u}_{2n}^{(2)} \mathbf{u}_{2n}^{(2)'} & \\ & & & \mathbf{u}_{2n}^{(2)} \mathbf{u}_{2n}^{(2)'} \otimes \mathbf{u}_{1n}^{(2)} \mathbf{u}_{1n}^{(2)'} \end{bmatrix} + \\ &+ \begin{bmatrix} 0 & & & \\ \mathbf{U}_{1s} \mathbf{U}'_{1s} & & & \\ & \mathbf{U}_{2s} \mathbf{U}'_{2s} & & \\ & & \mathbf{u}_{2n}^{(2)} \mathbf{u}_{2n}^{(2)'} \otimes \mathbf{U}_{1s} \mathbf{U}'_{1s} + \mathbf{U}_{2s} \mathbf{U}'_{2s} \otimes \mathbf{u}_{1n}^{(2)} \mathbf{u}_{1n}^{(2)'} + \mathbf{U}_{2s} \mathbf{U}'_{2s} \otimes \mathbf{U}_{1s} \mathbf{U}'_{1s} \end{bmatrix}. \end{aligned} \quad (4.33)$$

Let be \mathbf{U}_k , the orthogonal matrix of eigenvectors of the SVD of $\mathbf{D}'_k \mathbf{D}_k$, where \mathbf{U}_k is the partitioned matrix $\mathbf{U}_k = [\mathbf{U}_{kn} : \mathbf{U}_{ks}]$. We can take any eigenvectors, such that:

$$\mathbf{I}_{c_k} = \mathbf{U}_k \mathbf{U}'_k = \mathbf{U}_{kn} \mathbf{U}'_{kn} + \mathbf{U}_{ks} \mathbf{U}'_{ks}, \quad (4.34)$$

for $k = 1, 2$. Thus, for a second order penalty, we can take $\mathbf{U}_{kn} = [\mathbf{1}_k^* : \mathbf{u}_k^*]$, the $c_k \times 2$, matrix of eigenvectors, where $\mathbf{1}_k^* = \mathbf{1}_{c_k} / \sqrt{c_k}$, with $\mathbf{1}_k$ a vector of ones of length $c_k \times 1$ and \mathbf{u}_k^* is the vector $(1, 2, \dots, c_k)$ centered and scaled to have unit length. Then in (4.34), we can rewrite $\mathbf{U}_{kn} \mathbf{U}'_{kn} = \mathbf{1}_k^* \mathbf{1}_k^{*'} + \mathbf{u}_k^* \mathbf{u}_k^{*'}$, and thus

$$\mathbf{U}_{ks} \mathbf{U}'_{ks} = \mathbf{I}_{c_k} - \mathbf{1}_k^* \mathbf{1}_k^{*'} - \mathbf{u}_k^* \mathbf{u}_k^{*'} \quad (4.35)$$

where $\mathbf{1}_k^* \mathbf{1}_k^{*'} = \mathbf{1}_k \mathbf{1}_k' / c_k$. Then, if we replace in (4.33), the column vectors $\mathbf{u}_{kn}^{(2)}$ by \mathbf{u}_k^* , and given that

$$\mathbf{U}_{ks} \mathbf{U}_{ks}' + \mathbf{u}_k^* \mathbf{u}_k^{*'} = \mathbf{I}_{c_k} - \mathbf{1}_k \mathbf{1}_k' / c_k. \quad (4.36)$$

We obtain that $\mathbf{T}\mathbf{T}'$ is the matrix \mathbf{K} defined in (4.31).

To demonstrate that the matrix \mathbf{K} in (4.31) is a “contrast matrix” that centers the regression coefficient $\boldsymbol{\theta}$, we have that given the recovered penalty $\check{\mathbf{P}}$ defined in (4.30), the penalty over the original coefficients is $\boldsymbol{\theta}' \check{\mathbf{P}} \boldsymbol{\theta}$, that by 4.1, (i.e. $\check{\mathbf{P}} = \mathbf{K} \mathbf{P} \mathbf{K}$), can be written as:

$$\boldsymbol{\theta}' \check{\mathbf{P}} \boldsymbol{\theta} = \boldsymbol{\theta}' (\mathbf{K} \mathbf{P} \mathbf{K}) \boldsymbol{\theta} = \check{\boldsymbol{\theta}}' \mathbf{P} \check{\boldsymbol{\theta}},$$

where $\check{\boldsymbol{\theta}}$ are the centered regression coefficients, i.e.

$$\check{\boldsymbol{\theta}} = \mathbf{K} \boldsymbol{\theta} = \begin{pmatrix} 1 & \cdots & & \\ \vdots & \mathbf{K}_1 & & \\ & & \mathbf{K}_2 & \\ & & & (\mathbf{K}_2 \otimes \mathbf{K}_1) \end{pmatrix} \begin{pmatrix} \gamma \\ \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \boldsymbol{\theta}_s \end{pmatrix} = (\gamma, \check{\boldsymbol{\theta}}_1, \check{\boldsymbol{\theta}}_2, \check{\boldsymbol{\theta}}_s)',$$

and where:

$$\check{\boldsymbol{\theta}}_1 = \mathbf{K}_1 \boldsymbol{\theta}_1, \quad (4.37)$$

$$\check{\boldsymbol{\theta}}_2 = \mathbf{K}_2 \boldsymbol{\theta}_2, \quad (4.38)$$

$$\check{\boldsymbol{\theta}}_s = (\mathbf{K}_2 \otimes \mathbf{K}_1) \boldsymbol{\theta}_s. \quad (4.39)$$

The results in (4.37) and (4.38), center the regression coefficients for the additive main effects $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Equation (4.39), can be rewritten in array form as $\mathbf{K}_1 \boldsymbol{\Theta}_s \mathbf{K}_2$, and so, we are centering the coefficients matrix $\boldsymbol{\Theta}_s$ by rows and columns. ■

Theorem 4.3 (Linear constraints in two-dimensional S-ANOVA model regression coefficients). *The reparameterization of the S-ANOVA in (4.18) using the mixed model approach applies constraints over the regression coefficients $\boldsymbol{\theta}$, which are exactly equivalent to those applied in a factorial design with two main effects and a 2-way interaction, i.e.:*

$$\sum_i^{c_1} \boldsymbol{\theta}_{1i} = \sum_j^{c_2} \boldsymbol{\theta}_{2j} = 0, \quad \text{for main effects and} \quad (4.40)$$

$$\sum_i^{c_1} \boldsymbol{\Theta}_{ij} = \sum_j^{c_2} \boldsymbol{\Theta}_{ij} = 0, \quad \text{for 2-way interactions.} \quad (4.41)$$

Proof. It follows from 4.3. The centered regression coefficients defined in (4.37), (4.38) and (4.39), imply that the sum of this coefficients are zero. ■

This methodology can be extended for more dimensions. In next Section, we will discuss the case of $k = 3$, where all main effects and interactions are incorporated in the modelling. We illustrate how S-ANOVA model matrices and penalty can be directly constructed without explicit construction of the transformation matrix.

4.2.4 Smooth-ANOVA models construction

In previous sections, we have shown how to construct identifiable additive models with interactions using a reparameterization into a mixed model. In practice, it is not necessary to construct the transformation matrix T in order to obtain the fixed and random effects matrices. The definition of the matrix T , used to demonstrate that the model construction procedure, yields the restrictions on the regression coefficients in the non-transformed P -spline model. In this Section, we take advantage of the methodology to build ANOVA-type models easily, i.e. given the functional form of the S-ANOVA model, we can construct directly the model matrices X and Z , and then the block-diagonal penalty matrix F .

For example, let us consider ANOVA-type decomposition of three covariates, with terms:

$$\begin{aligned} y = & \gamma + f_1(x_1) + f_2(x_2) + f_3(x_3) + \\ & + f_{1,2}(x_1, x_2) + f_{1,3}(x_1, x_3) + f_{2,3}(x_2, x_3) + \\ & + f_{1,2,3}(x_1, x_2, x_3) + \epsilon . \end{aligned} \quad (4.42)$$

Model (4.42) is the *full* S-ANOVA model which includes all main effects, 2-way and 3-way interactions, we can proceed as in the two-dimensional case in Section 4.2.1, and demonstrate that the constraints on the coefficients are those in Table 4.1.

The S-ANOVA model in (4.42) can be constructed term by term. In previous sections, we have already seen each of the smooth terms independently, and therefore, we can construct the model matrices (we assume a second order penalty):

- For the main effects:

$$f_k(x_k) = [x_k : Z_k], \text{ for } k = 1, 2, 3.$$

Table 4.1: Set of regression coefficient constraints in a full 3d P -spline ANOVA-type model.

Constraints	
Main effects	$\sum_i^{c_1} \theta_i^{(1)} = \sum_j^{c_2} \theta_j^{(2)} = \sum_k^{c_3} \theta_k^{(3)} = 0$
2-way interaction	$\sum_i^{c_1} \theta_{ij}^{(1,2)} = \sum_j^{c_2} \theta_{ij}^{(1,2)} = \sum_i^{c_1} \theta_{ik}^{(1,3)} = \sum_k^{c_3} \theta_{ik}^{(1,3)} = \sum_j^{c_2} \theta_{jk}^{(2,3)} = \sum_k^{c_3} \theta_{jk}^{(2,3)} = 0$
3-way interaction	$\sum_i^{c_1} \theta_{ijk}^{(1,2,3)} = \sum_j^{c_2} \theta_{ijk}^{(1,2,3)} = \sum_k^{c_3} \theta_{ijk}^{(1,2,3)} = 0$

- For two-way interactions, we have:

$$\begin{aligned}
 f_{i,j}(\mathbf{x}_i, \mathbf{x}_j) &= \bigotimes_{i < j}^k [\mathbf{x}_i : \mathbf{Z}_i] = [\mathbf{x}_i : \mathbf{Z}_i] \otimes [\mathbf{x}_j : \mathbf{Z}_j] = \\
 &\equiv [\mathbf{x}_i \otimes \mathbf{x}_j : \mathbf{x}_i \otimes \mathbf{Z}_j : \mathbf{Z}_i \otimes \mathbf{x}_j : \mathbf{Z}_i \otimes \mathbf{Z}_j].
 \end{aligned}$$

- For the three-way interaction:

$$\begin{aligned}
 f_{1,2,3}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) &= \bigotimes_{i=1}^k [\mathbf{x}_i : \mathbf{Z}_i] = [\mathbf{x}_1 : \mathbf{Z}_1] \otimes [\mathbf{x}_2 : \mathbf{Z}_2] \otimes [\mathbf{x}_3 : \mathbf{Z}_3] = \\
 &\equiv [\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3 : \mathbf{Z}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3 : \mathbf{x}_1 \otimes \mathbf{Z}_2 \otimes \mathbf{x}_3 : \mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{Z}_3 : \\
 &\quad \mathbf{Z}_1 \otimes \mathbf{Z}_2 \otimes \mathbf{x}_3 : \mathbf{Z}_1 \otimes \mathbf{x}_2 \otimes \mathbf{Z}_3 : \mathbf{x}_1 \otimes \mathbf{Z}_2 \otimes \mathbf{Z}_3 : \mathbf{Z}_1 \otimes \mathbf{Z}_2 \otimes \mathbf{Z}_3].
 \end{aligned}$$

Given that only the random part is penalized, we can construct the block-diagonal penalty \mathbf{F} , for each smooth term:

- For each main effect:

$$\mathbf{F}_k = \lambda_k \tilde{\Sigma}_k, \text{ for } k = 1, 2, 3.$$

- For each two-way interaction:

$$\mathbf{F}_{i,j} = \begin{bmatrix} \tau_i \tilde{\Sigma}_i & & \\ & \tau_j \tilde{\Sigma}_j & \\ & & \tau_j \tilde{\Sigma}_j \oplus \tau_i \tilde{\Sigma}_i \end{bmatrix}, \text{ for } i < j.$$

- For the three-way interaction:

$$\mathbf{F}_{1,2,3} = \begin{bmatrix} \bigoplus_k^3 \phi_k \tilde{\Sigma}_k \\ \bigoplus_{i < j}^3 (\phi_i \tilde{\Sigma}_i \oplus \phi_j \tilde{\Sigma}_j) \\ \phi_1 \tilde{\Sigma}_1 \oplus \phi_2 \tilde{\Sigma}_2 \oplus \phi_3 \tilde{\Sigma}_3 \end{bmatrix}, \text{ for } i < j \text{ and } k = 1, 2, 3.$$

Note that, model matrices \mathbf{X} and \mathbf{Z} are exactly the same as those obtained in a three-dimensional model in (2.79) and (2.80), but with elements reordered according to the S-ANOVA model formulation. The penalty \mathbf{F} for this model, has seven blocks, as the penalty in the three-dimensional model in (2.81), but in the S-ANOVA model, each smooth function is penalized independently with smoothing parameters: λ_k , for the $k = 1, 2, 3$ main effects, τ_1, τ_2 , and τ_3 for the two-way interactions, and ϕ_1, ϕ_2 and ϕ_3 for the three-way interaction. In other words, we allow for a more flexible model through imposing independent and separate penalizations, and considering a different amount of smoothing for each smooth function.

An interesting question is the suitability of the S-ANOVA models in comparison to an additive model or an interaction model. We might be interesting in checking if the interaction term is significative or not. In Section 4.2.5 we analyze the performance of the S-ANOVA model in a simulation study with different scenarios. We will discuss the problem of identifying the smooth terms in these type of models in Section 4.3.

4.2.5 Simulation of smooth surfaces

In this section we examine the performance of the S-ANOVA model in comparison to additive and interaction models. For simplicity, we simulated $\boldsymbol{\eta}$ as a function of two covariates \mathbf{x}_1 and \mathbf{x}_2 , in three different scenarios:

$$\boldsymbol{\eta}^{(1)} = f_1(\mathbf{x}_1) + f_1(\mathbf{x}_2), \quad (\text{"Two main effects model"})$$

$$\boldsymbol{\eta}^{(2)} = f_{1,2}(\mathbf{x}_1, \mathbf{x}_2), \text{ and} \quad (\text{"Interaction model"})$$

$$\boldsymbol{\eta}^{(3)} = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + f_{1,2}(\mathbf{x}_1, \mathbf{x}_2). \quad (\text{"Two main effects with interaction"})$$

Each of the main effects smooth functions f_1 and f_2 are non-linear and the interaction function as a complex surface. The functions we have used are:

$$f_1(\mathbf{x}_1) = \sin(2\pi\mathbf{x}_1), \quad (4.43)$$

$$f_2(\mathbf{x}_2) = \cos(3\pi\mathbf{x}_2), \text{ and} \quad (4.44)$$

$$f_{1,2}(\mathbf{x}_1, \mathbf{x}_2) = 3 \sin(2\pi\mathbf{x}_1) (2\mathbf{x}_2 - 1). \quad (4.45)$$

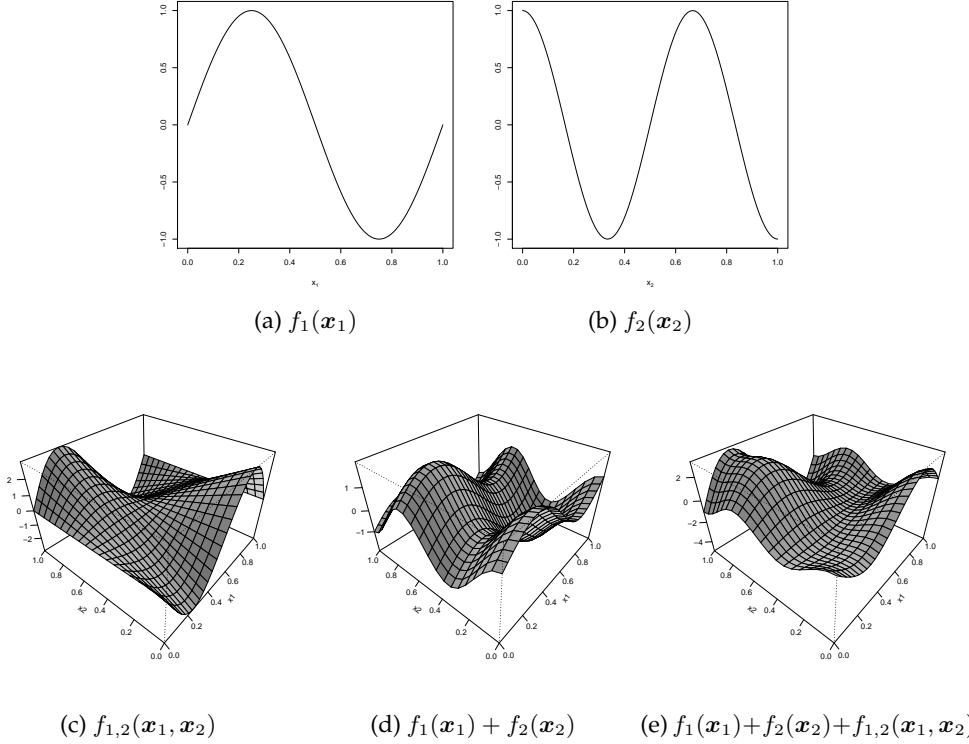


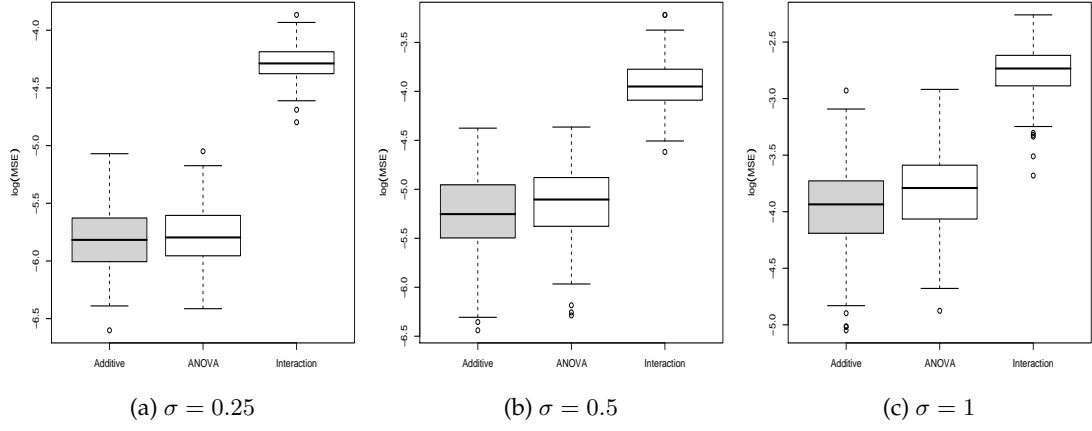
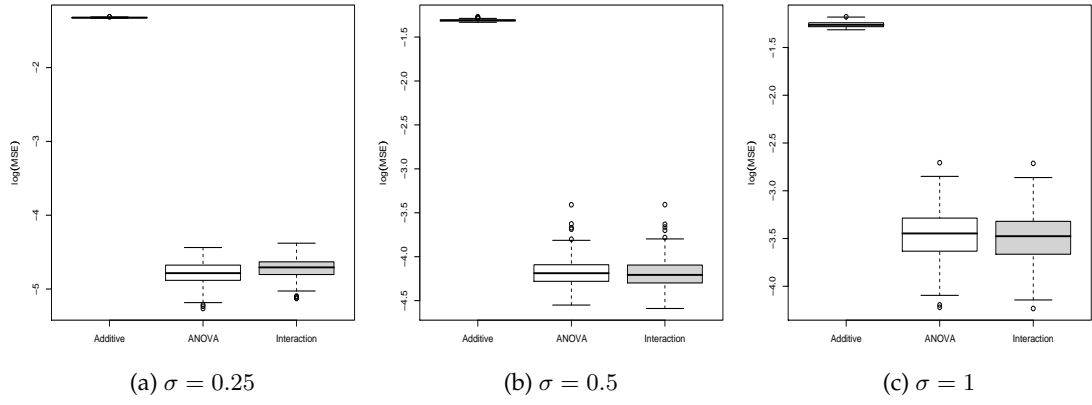
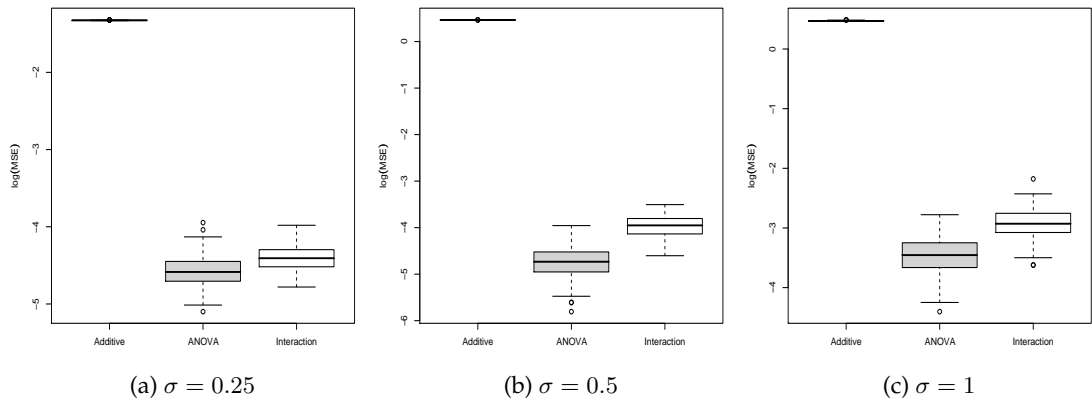
Figure 4.1: Simulated functions: (a) and (b) are the nonlinear main effects of x_1 and x_2 ; (c) is the additive surface of main effects; (d) is interaction surface and (e) is the sum of the main effects and the interaction surfaces.

We consider the case of data on a regular grid, the covariates values of x_1 and x_2 are chosen in the interval $[0, 1]$ in a regular pattern, with dimensions $n_1 = 30$ and $n_2 = 20$, respectively, we obtain a grid of $n = 600$ values. Figure 4.1 shows the simulated true smooth functions and true surfaces for the proposed scenarios.

For each scenario we fitted three smooth mixed models: *additive*, *anova* and *interaction* models. We constructed the marginal B -splines bases B_1 and B_2 with 8 and 6 knots respectively, with cubic splines and second order penalties. The models were estimated by REML. To check each model's performance we computed the mean square error for each replicate as:

$$\text{MSE}^{(i)} = \frac{1}{n} \sum_{i=1}^n (\eta^{(i)} - \hat{\eta}_r^{(i)})^2, \text{ for } i = 1, 2, 3 \text{ and } r = 1, \dots, R.$$

where $\eta^{(i)}$ is the true simulated surface under each i^{th} scenario and $\hat{\eta}^{(r)}$ is the estimated function for each model at each $r = 1, \dots, R$ replicates. Figure 4.4 shows the boxplots of

Figure 4.2: $\log(\text{MSE})$ of fitted smooth models in scenario 1 and $R = 200$.Figure 4.3: $\log(\text{MSE})$ of fitted smooth models in scenario 2 and $R = 200$.Figure 4.4: $\log(\text{MSE})$ of fitted smooth models in scenario 3 and $R = 200$.

the $\log(\text{MSE})$ values for fitted smooth models and for each of the simulated scenarios with $\sigma = \{0.25, 0.5, 1\}$ and $R = 200$. The grey shaded boxplot corresponds to the model from which we have simulated each scenario, i.e. in scenario 1, we consider $\eta^{(1)}$ as a function of two main effects, and thus the *additive* model is the favoured model, and we compare *anova* and *interaction* models performance; in scenario 2, $\eta^{(2)}$ is an interaction surface, thus the favoured model is the *interaction* model; and in scenario $\eta^{(3)}$, we have simulated from two main effects with and interaction, thus the *anova* model is the favoured model.

The results of the simulation study for each scenario are summarized as follows:

- In the first scenario, the true surface, $\eta^{(1)}$, is constructed from two main effects. Thus, the *additive* model fit is the most adequate. The *anova* model, in this scenario is reduced to the *additive* model with smoothing parameters for the interaction, τ_1 and τ_2 that tend to ∞ . This reflects that the interaction penalty is ineffective. Note that, since $\tau_1, \tau_2 \rightarrow \infty$, there might be some small numerical differences in the estimation of the models, since we are estimating the smoothing parameters by REML, and therefore the boxplots of *additive* and *anova* models are not exactly similar in a few replicates. We have considered an upper bound for the smoothing parameters equal to 10^6 .
- In the second scenario, the true surface $\eta^{(2)}$, is purely interaction. As shown in [Figure 4.3](#), the *additive* model is the worst in terms of accuracy (higher MSE values), since it does not account for modelling the interaction. The *anova* model has the same performance as the *interaction* model, and the smoothing parameters for the additive part, λ_1 , and λ_2 will tend to ∞ (i.e. no additive penalty effect). The interaction term for the *anova* model will capture the interaction effect. As noticed in the previous case, there might be small numerical differences between *anova* and *interaction* models, due to the fact that $(\lambda_1, \lambda_2) \rightarrow \infty$.
- Finally, in the third scenario. We simulated a true surface, $\eta^{(3)}$, with two main effects with an interaction. The worst performance corresponds to the *additive* fit, that is constrained to model the true model with an additive formulation. The best model performance is the *anova*.

Conclusions of the simulation study

We have performed a small simulation study to analyze the performance of the S-ANOVA model. We conclude that S-ANOVA performs as well as the most adequate model in all of the three scenarios proposed. Given the construction of the S-ANOVA model bases

and penalty, each term is identifiable and captures each of the simulated true functions. A question of statistical interest that arises from this study is the model selection and testing for smooth terms. Notice that, reformulating a P -spline model as a mixed model, the smoothing parameter is the ratio of two variances, i.e., $\lambda = \sigma^2/\sigma_\alpha^2$, where σ_α^2 is the variance of the random effect α .

In the context of the S-ANOVA models, if we consider the simulation scenarios in this Section, the interest lies in testing for the variance components of the random effects. For instance, in scenario 1, we could be interested in testing the additive model *versus* the S-ANOVA, that means testing if the smoothing parameters for the interaction τ_1 and τ_2 tend to ∞ , or in terms of the variance components if the variances of the interaction random effects: $\sigma_{\tau_1}^2$ and $\sigma_{\tau_2}^2$ are equal to zero or not. Another possibility is to test the adequacy of a S-ANOVA model *versus* an interaction model, i.e. if $\lambda_1, \lambda_2 \rightarrow \infty$, or equivalently if $\sigma_{\lambda_1}^2 = \sigma_{\lambda_2}^2 = 0$.

The general problem of testing for interaction terms is not trivial, in the next Section we present a brief literature review of the methods and present the difficulties of implementing the existing methods to the S-ANOVA models we propose in this Chapter.

4.3 Testing components in smoothing mixed models

In recent years, many authors have paid attention in the theoretical aspects of the P -spline methodology (Hall and Opsomer, 2005; Claeskens et al., 2009). From the mixed model point of view, and as GAMs (Wand, 1999; Aerts et al., 2002; Käüermann, 2005; Käüermann et al., 2009). These results have been the starting point for some asymptotic and inferential aspects of the P -spline methodology. A question of interest is to develop formal statistical tests for the functional forms of smoothing models. Several methods have been proposed for testing in the context of smoothing models. Hastie and Tibshirani (1990) show how the residual sum of squares for competing models can be compared using approximate degrees of freedom, by analogy to the F -type test in linear models, in the context of additive models (Cantoni and Hastie, 2000). Bowman and Azzalini (1997) presents how the p -values can be efficiently computed using quadratic forms.

However, in the context of P -splines these tests do not take into account the uncertainty in the estimation of the smoothing parameters. When P -splines are formulated as mixed models, testing the presence of a smooth term is equivalent to test if the corresponding variance component for the random effect is zero. In this context, the usual test for random effects variance components involves the (restricted) likelihood ratio

tests statistics (RLRT), defined as:

$$\text{RLRT} = \sup_{\lambda \in H_1} \mathcal{L}_R(\lambda, \sigma^2) - \sup_{\lambda \in H_0} \mathcal{L}_R(\lambda, \sigma^2), \quad (4.46)$$

where \mathcal{L}_R is the residual log-likelihood in (2.32) (Pinheiro and Bates, 2000; Verbeke and Molenberghs, 2000). Given that REML uses the likelihood of residuals after fitting the fixed effects, models fitted with different fixed effects structures cannot be compared on the basis of their restricted likelihoods. Self and Liang (1987, 1995) and Stram and Lee (1994), discussed the asymptotic distribution of RLRT statistic and showed that, under the assumption that the response vector \mathbf{y} can be partitioned into independent subvectors, and the number of subvectors tends to infinity, RLRT has a $\frac{1}{2}\chi_q + \frac{1}{2}\chi_{q+1}$ asymptotic distribution, where q is the number of fixed effects under the null hypothesis. However, (4.46) may not be appropriate under the alternative hypothesis in some type of models, leading to a poor approximation by the Chi-square mixture. Crainiceanu and Ruppert (2004) suggest the use of simulations to determine the null distribution of the (restricted) likelihood ratio test statistic. The idea is to estimate the model parameters under the null hypothesis, then simulate the distribution of the (restricted) likelihood ratio test under the null model at the parameters. Let us suppose an univariate P -spline model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\alpha + \epsilon, \text{ with } \alpha \sim \mathcal{N}(0, \sigma_\alpha^2 \mathbf{I}), \text{ and } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Testing for such absence of random effect has to take into account that the tested parameter is on the boundary of the parameter space. If we are interested in testing the presence of a smooth term (parametric *versus* non-parametric), we will consider:

$$H_0 : \sigma_\alpha^2 = 0 \quad \text{versus} \quad H_1 : \sigma_\alpha^2 > 0.$$

For testing one variance components, Crainiceanu and Ruppert (2004) derive the finite sample and asymptotic distribution of the (restricted) likelihood test statistic. They proposed an efficient simulation algorithms of their null distributions, based on the spectral decomposition of the likelihood ratio tests. For the special case of testing all variance components simultaneously, i.e., for an additive model with L random effects:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\alpha_1 + \mathbf{Z}_2\alpha_2 + \dots + \mathbf{Z}_L\alpha_L + \epsilon,$$

where $(\alpha_1, \dots, \alpha_L)' \sim \mathcal{N}(0, \text{blockdiag}(\sigma_{\alpha_1}^2 \mathbf{I}, \dots, \sigma_{\alpha_L}^2 \mathbf{I}))$. Claeskens (2004) provides a the spectral representation of the restricted log-likelihood of a model with L variance components. Thus, the test is $H_0 : \sigma_{\alpha_1}^2 = \sigma_{\alpha_2}^2 = \dots = \sigma_{\alpha_L}^2 = 0$. However, the algorithm becomes computationally intensive and nearly impractical. Given these limitations

Crainiceanu and Ruppert (2004), conclude that the use of a parametric bootstrap instead may be a good strategy for the general case. The recent work by Greven et al. (2008) proposed two alternative approximations to the null distribution for testing particular cases for more than one variance components that avoids parametric bootstrap. Scheipl et al. (2008) presented several simulation studies for testing zero variance components and have implemented these methods in the R package `RLRsim`.

In the context of the S-ANOVA models presented in this Chapter, the development of tests based on RLRT statistics are not completely possible to implement using the methodology proposed in Greven et al. (2008), unless we consider an isotropic interaction (i.e. a single smoothing parameter, τ , for the interaction). Consider the problem of testing a smooth additive mixed model with two covariates \mathbf{x}_1 and \mathbf{x}_2 , against a S-ANOVA model, i.e.:

$$H_0 : f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) \quad \text{versus} \quad H_1 : f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + f_{1,2}(\mathbf{x}_1, \mathbf{x}_2), \quad (4.47)$$

where the model under the alternative is the S-ANOVA in (4.18), that in mixed model formulation is written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \text{with } \boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{G}) \text{ and } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2),$$

where the fixed effects matrix \mathbf{X} must be equal to the fixed effects matrix of the *null* additive model, without the interaction $\mathbf{x}_2 \otimes \mathbf{x}_1$. The random effects matrix is the same as defined in (4.26), with variance-covariance matrix for the random effects \mathbf{G} , given by $\mathbf{G} = \sigma^2 \mathbf{F}^{-1}$. Then, considering an isotropic penalty for the interaction, the block-diagonal mixed model penalty, \mathbf{F} , is defined as:

$$\mathbf{F} = \begin{pmatrix} \lambda_1 \tilde{\boldsymbol{\Sigma}}_1 & & & & \\ & \lambda_2 \tilde{\boldsymbol{\Sigma}}_2 & & & \\ & & \tau \tilde{\boldsymbol{\Sigma}}_1 & & \\ & & & \tau \tilde{\boldsymbol{\Sigma}}_2 & \\ & & & & \tau \mathbf{I}_{c_2-2} \otimes \tilde{\boldsymbol{\Sigma}}_1 + \tau \tilde{\boldsymbol{\Sigma}}_2 \mathbf{I}_{c_1-2} \end{pmatrix}. \quad (4.48)$$

Thus, testing for a smooth interaction is equivalent to test if the smoothing parameter τ tends to ∞ , or equivalently, in terms of variance components, given that $\tau = \sigma^2 / \sigma_\tau^2$, we test:

$$H_0 : \sigma_\tau^2 = 0 \quad \text{versus} \quad H_1 : \sigma_\tau^2 > 0, \quad (4.49)$$

where σ_τ^2 is the variance of random effect for the isotropic interaction. This test is only as a particular case of interaction testing and addresses the difficulties of applying the

existing methods to the S-ANOVA model we propose in this Chapter.

The testing problem is a current topic of research in P -splines smoothing as mixed models, that combines theoretical and computational aspects of interest. For the general case of the S-ANOVA models we have developed in this Chapter, the incorporation of several covariates and its anisotropic interactions increases the complexity and the applicability of the restricted log-likelihood approaches already developed.

4.4 Reduced S-ANOVA models for spatio-temporal data

In some real applications, it might be interesting to include only some smooth terms and ignore others, i.e. we can choose which main effects or interactions to include as part of the model. This will lead us to a more flexible and interpretable multidimensional smooth model than an additive or interaction model. We call these models: *reduced* S-ANOVA models, since it does not incorporate all the components of the full ANOVA-type decomposition. For example, instead of considering the full or complete ANOVA-type model in (4.42), we might be interested in a S-ANOVA model of the form:

$$\mathbb{E}[\mathbf{y}] = \gamma + f_1(\mathbf{x}_1) + f_{2,3}(\mathbf{x}_2, \mathbf{x}_3) + f_{1,2,3}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3), \quad (4.50)$$

with one main effect and only a two-way interaction, and the three-way interaction.

The spatio-temporal data smoothing is an interesting application in which a model of form (4.50) might result very useful and easy to interpret. In many cases, the main interest when considering spatio-temporal data is to explicitly model the space-time interaction, since a separable model of an additive function of space and a temporal trend, does not reflect the real underlying process.

In Lee and Durbán (2010) we applied the S-ANOVA methodology to the analysis of spatio-temporal data. We propose a new model based on P -splines for spatio-temporal smoothing, using a reduced S-ANOVA decomposition as in (4.50), where the model includes a bivariate spatial smooth term, a one-dimensional smooth term for the time trend, and a three-dimensional smooth term for the space-time interaction. We will follow the methodology developed in Section 4.2 to demonstrate which are the constraints on the regression coefficients using our reparameterization. As a reduced model, we will show that these linear constraints will include only a subset of the restrictions of the full ANOVA model in shown in Table 4.1. In Section 4.4.4, we apply the methodology to air pollution of ozone levels in Europe during 1999-2005.

4.4.1 Spatio-temporal P -spline models and basis

In spatio-temporal data, the response variable \mathbf{y} , is usually measured in scattered/spatial locations but also at regular time intervals. This yields a three-dimensional model of spatial coordinates x_1, x_2 and time x_t . We start by considering an interaction model with functional form given by:

$$\mathbb{E}[\mathbf{y}] = f_{s,t}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_t). \quad (4.51)$$

Model (4.51) is a three-dimensional function across space and time, and explicitly considers the space-time interaction (a non-separable model). From a multidimensional P -spline approach, we propose the model (4.51) with a regression basis consisting of the Kronecker product of two basis functions: (i) a basis for smoothing over the spatial surface, and (ii) a basis for smoothing over the temporal dimension. This leads us to a spatio-temporal B -spline basis given by:

$$\mathbf{B} = \mathbf{B}_s \otimes \mathbf{B}_t, \quad nt \times c_s c_t, \quad (4.52)$$

where \mathbf{B}_s is the spatial B -spline basis as we defined in Chapter 3 (i.e. $\mathbf{B}_2 \square \mathbf{B}_1$), of dimension $n \times c_s$, where $c_s = c_1 c_2$, and \mathbf{B}_t is the marginal B -spline basis for time, of dimension $t \times c_t$. Note that, in model (4.51), we may replace the $nt \times 1$ response vector \mathbf{y} , by the matrix \mathbf{Y} of dimension $t \times n$, and the coefficient vector $\boldsymbol{\theta}$ of length $c_s c_t \times 1$, by an array of coefficients $\boldsymbol{\Theta}$, of dimension $c_t \times c_s$. In matrix notation, we can rewrite the model as:

$$\mathbb{E}[\mathbf{Y}] = \mathbf{B}_t \boldsymbol{\Theta} \mathbf{B}_s'. \quad (4.53)$$

Therefore, model (4.53), can be considered as a GLAM of space and time, and the array algorithms shown in Section 2.3 can be used to fit the model. We replace the $nt \times 1$ response vector \mathbf{y} , by the matrix \mathbf{Y} of dimension $t \times n$, and the coefficient vector $\boldsymbol{\theta}$ of length $c_s c_t \times 1$, by an array of coefficients $\boldsymbol{\Theta}$, of dimension $c_t \times c_s$.

As we are considering a three-dimensional model, smoothness is imposed via a penalty matrix \mathbf{P} in (2.77). This penalty (2.77) allows spatial anisotropy with smoothing parameters λ_1 and λ_2 for the spatial coordinates, and a smoothing parameter λ_t , for the temporal component. We use the mixed model reparameterization of the three-dimensional P -spline model in (4.51). In this case, note that, for the spatial component we use the results shown in Chapter 3, for the row Tensor or Box-product. The new

bases for model (4.51) can be written in a compact notation as:

$$\mathbf{X} = \mathbf{X}_s \otimes \mathbf{X}_t, \text{ and} \quad (4.54)$$

$$\mathbf{Z} = [\mathbf{Z}_s \otimes \mathbf{X}_t : \mathbf{X}_s \otimes \mathbf{Z}_t : \mathbf{Z}_s \otimes \mathbf{Z}_t], \quad (4.55)$$

where \mathbf{X}_s and \mathbf{Z}_s are the matrices defined in (3.3) and (3.4) for the spatial case (and involves the row Tensor product). And \mathbf{X}_t and \mathbf{Z}_t are the fixed and random effects matrices for the temporal dimension. The block-diagonal mixed model penalty \mathbf{F} , is the same as the one obtain in (2.81), with seven blocks:

$$\mathbf{F} = \text{blockdiag}(\mathbf{F}_{(1)}, \mathbf{F}_{(2)}, \mathbf{F}_{(1,2)}, \mathbf{F}_{(t)}, \mathbf{F}_{(1,t)}, \mathbf{F}_{(2,t)}, \mathbf{F}_{(1,2,t)}).$$

As we showed in Section 2.3.2, the construction of the new bases (4.54) and (4.55) allows us to represent the fitted values in terms of the sum of additive components plus interactions (2-way and 3-way interactions). For spatio-temporal data, this decomposition may be very useful in terms of the interpretability of the model fit, since we can decompose the overall fit not only as main effects of latitude and longitude, (or other covariates), but also the spatial effects (2-way interaction), and specially the interaction between space and time (3-way interactions). However, in terms of model formulation, it does not account for independent and separate penalties since we have three smoothing parameters λ_1, λ_2 and λ_t for each of the dimensions of the model. That is, the amount of smoothing used for the additive terms is also used for the interactions. In some cases (as we will show in the analysis of the ozone data), this is not realistic, and so, we will apply the P -spline ANOVA methodology to the spatio-temporal setting.

We use a reduced S-ANOVA model of the form:

$$\mathbf{y} = \gamma + f_s(\mathbf{x}_1, \mathbf{x}_2) + f_t(\mathbf{x}_t) + f_{st}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_t) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad (4.56)$$

where we explicitly consider a smooth term for the spatial surface, for temporal smooth trend, and a smooth term for space-time interaction. A B -spline regression basis for model (4.56) would be:

$$\mathbf{B} = [\mathbf{1}_{nt} : \mathbf{B}_s \otimes \mathbf{1}_t : \mathbf{1}_n \otimes \mathbf{B}_t : \mathbf{B}_s \otimes \mathbf{B}_t], \quad (4.57)$$

with vector of regression coefficients: $\boldsymbol{\theta} = (\gamma, \boldsymbol{\theta}^{(s)'}, \boldsymbol{\theta}^{(t)'}, \boldsymbol{\theta}^{(st)'})'$. The penalty matrix is block-diagonal with penalties over $\boldsymbol{\theta}$ of the form:

$$\mathbf{P} = \text{blockdiag}(0, \mathbf{P}_{(s)}, \mathbf{P}_{(t)}, \mathbf{P}_{(st)}), \quad (4.58)$$

where $P_{(s)}$ is the two-dimensional penalty matrix for the spatial component, with smoothing parameters λ_1 and λ_2 as in (2.63), i.e.

$$P_{(s)} = \lambda_1 I_{c_2} \otimes D_1' D_1 + \lambda_2 D_2' D_2 \otimes I_{c_2}, \quad (4.59)$$

$P_{(t)}$ is the one-dimensional penalty matrix for the time component, with smoothing parameter λ_t , given by:

$$P_{(t)} = \lambda_t D_t' D_t, \quad (4.60)$$

and $P_{(st)}$ is the three-dimensional penalty matrix for the spatio-temporal component as in (2.77), with smoothing parameters τ_1 , τ_2 and τ_t :

$$P_{(st)} = \tau_2 D_2' D_2 \otimes I_{c_1} \otimes I_{c_t} + \tau_1 I_{c_2} \otimes D_1' D_1 \otimes I_{c_t} + \tau_3 I_{c_2} \otimes I_{c_1} \otimes D_t' D_t. \quad (4.61)$$

The reduced S-ANOVA model in (4.56), include univariate, bivariate and three-dimensional terms. We have already seen the components of each smooth function in previous chapters, and thus, we are able to construct the mixed model bases for each smooth term and remove the repeated terms, as we showed in Sections 4.2.1 and 4.2.4, without considering the transformation matrix T , such that $BT = [X : Z]$. For basis in (4.57), we define the mixed model matrices:

$$\begin{aligned} X &= [\underbrace{X_s \otimes \mathbf{1}_t}_{f_s(\mathbf{x}_1, \mathbf{x}_2)} : \underbrace{\mathbf{1}_n \otimes \mathbf{x}_t}_{f_t(\mathbf{x}_t)} : \underbrace{\mathbf{x}_s \otimes \mathbf{x}_t}_{f_{s,t}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_t)}] \\ Z &= [Z_s \otimes \mathbf{1}_t : \mathbf{1}_n \otimes Z_t : Z_s \otimes \mathbf{x}_t : \underbrace{[\mathbf{1}_n \square \mathbf{x}_1]}_{\mathbf{x}_1} : \underbrace{[\mathbf{x}_2 \square \mathbf{1}_n]}_{\mathbf{x}_2} : \mathbf{x}_s \otimes Z_t : Z_s \otimes Z_t], \end{aligned} \quad (4.62)$$

where $\mathbf{x}_s = \mathbf{x}_2 \square \mathbf{x}_1$. Both matrices are exactly the same as those defined in (4.54) and (4.55) for the spatio-temporal interaction model (4.51) but with a different order in the columns and blocks. The block-diagonal penalty F can be easily obtained. However, in order to demonstrate which are the linear constraints over the regression coefficients, we will construct the transformation matrix and obtain the penalty matrix F .

4.4.2 Transformation matrix in the reduced spatio-temporal S-ANOVA model

In order to reparameterize the model basis and coefficients, we must define the transformation matrix T that avoids the identifiability problem.

Proposition 4.4. *The transformation matrix T , such that, the reduced spatio-temporal S-ANOVA model in (4.56), is reparameterized into a mixed model and the identifiability problem*

is avoided, is defined as the partitioned matrix: $\mathbf{T} = [\mathbf{T}_n : \mathbf{T}_s]$, where we consider each block as:

$$\mathbf{T}_n = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \mathbf{T}_n^{(s)} & \\ & \mathbf{T}_n^{(t)} & \\ 0 & & \mathbf{T}_n^{(st)} \end{bmatrix} \quad \text{and} \quad \mathbf{T}_s = \begin{bmatrix} 0 & \cdots \\ \mathbf{T}_s^{(s)} & \\ \vdots & \mathbf{T}_s^{(t)} \\ & \mathbf{T}_s^{(st)} \end{bmatrix}, \quad (4.63)$$

where the first entry of 1 corresponds to the constant term.

Remark 4.5. We use the superscripts $^{(s)}$, $^{(t)}$ and $^{(st)}$, to denote that each sub-block corresponds to the spatial, temporal and spatio-temporal components of the decomposition respectively.

Proof of 4.4. As we showed in Section 4.2.1, removing the column vectors of ones in the model matrices is equivalent to remove the first columns in the null space eigenvectors of the SVD decomposition, and leave the vectors \mathbf{u}_1^* , \mathbf{u}_2^* and \mathbf{u}_t^* . However, given that in the reduced model, only some terms of the full ANOVA model are omitted, we must be care of which of these columns we have to remove. Let us define the B -spline basis for the reduced S-ANOVA model in (4.57) as:

$$\mathbf{B} = [\mathbf{1}_{nt} : \underbrace{\mathbf{B}_s \otimes \mathbf{1}_t}_{\mathbf{B}^{(s)}} : \underbrace{\mathbf{1}_n \otimes \mathbf{B}_t}_{\mathbf{B}^{(t)}} : \underbrace{\mathbf{B}_s \otimes \mathbf{B}_t}_{\mathbf{B}^{(st)}}], \quad \text{where } \mathbf{B}_s = \mathbf{B}_2 \square \mathbf{B}_1.$$

We consider each sub-block separately:

- For the spatial component, we define the sub-matrices:

$$\mathbf{T}_n^{(s)} = [\mathbf{U}_{2n} \otimes \mathbf{U}_{1n}] \otimes \mathbf{1}, \quad \text{and} \quad \mathbf{T}_s^{(s)} = [\mathbf{U}_{2s} \otimes \mathbf{U}_{1n} : \mathbf{U}_{2n} \otimes \mathbf{U}_{1n} : \mathbf{U}_{2s} \otimes \mathbf{U}_{1s}] \otimes \mathbf{1}.$$

Then, the fixed effects matrix for the spatial component is:

$$\begin{aligned} \mathbf{X}^{(s)} &= \mathbf{B}^{(s)} \mathbf{T}_n^{(s)} = (\mathbf{B}_s \otimes \mathbf{1}_t)([\mathbf{U}_{2n} \otimes \mathbf{U}_{1n}] \otimes \mathbf{1}) = (\mathbf{B}_2 \mathbf{U}_{2n} \square \mathbf{B}_1 \mathbf{U}_{1n}) \otimes \mathbf{1}_t = \\ &= (\mathbf{X}_2 \square \mathbf{X}_1) \otimes \mathbf{1}_t = \mathbf{X}_s \otimes \mathbf{1}_t, \end{aligned}$$

where $\mathbf{X}_1 = [\mathbf{1}_n : \mathbf{x}_1]$ and $\mathbf{X}_2 = [\mathbf{1}_n : \mathbf{x}_2]$. And the spatial random effects matrix is:

$$\begin{aligned} \mathbf{Z}^{(s)} &= \mathbf{B}^{(s)} \mathbf{T}_s^{(s)} = (\mathbf{B}_s \otimes \mathbf{1}_t)([\mathbf{U}_{2s} \otimes \mathbf{U}_{1n} : \mathbf{U}_{2n} \otimes \mathbf{U}_{1n} : \mathbf{U}_{2s} \otimes \mathbf{U}_{1s}] \otimes \mathbf{1}) = \\ &= (\mathbf{B}_2 \mathbf{U}_{2s} \square \mathbf{B}_1 \mathbf{U}_{1n} : \mathbf{B}_2 \mathbf{U}_{2n} \square \mathbf{B}_1 \mathbf{U}_{1s} : \mathbf{B}_2 \mathbf{U}_{2s} \square \mathbf{B}_1 \mathbf{U}_{1s}) \otimes \mathbf{1}_t = \\ &= (\mathbf{Z}_2 \square \mathbf{X}_1 : \mathbf{X}_2 \square \mathbf{Z}_1 : \mathbf{Z}_2 \square \mathbf{Z}_1) \otimes \mathbf{1}_t, \end{aligned}$$

where $Z_1 = B_1 U_{1s}$ and $Z_2 = B_2 U_{2s}$.

- For the temporal component, we take: $T_n^{(t)} = 1 \otimes \mathbf{u}_t^*$, and $T_s^{(t)} = 1 \otimes U_{ts}$. Then, the fixed effects matrix for the temporal component is:

$$\begin{aligned} X^{(t)} &= B^{(t)} T_n^{(t)} = (\mathbf{1}_n \otimes B_t)(1 \otimes \mathbf{u}_t^*) = \\ &= \mathbf{1}_t \otimes B_t \mathbf{u}_t^* = \mathbf{1}_t \otimes \mathbf{x}_t, \end{aligned}$$

and temporal random effects matrix is:

$$\begin{aligned} Z^{(t)} &= B^{(t)} T_s^{(t)} = (\mathbf{1}_n \otimes B_t)(1 \otimes U_{ts}) = \\ &= \mathbf{1}_t \otimes B_t U_{ts} = \mathbf{1}_t \otimes Z_t. \end{aligned}$$

where $Z_t = B_t U_{ts}$.

- Finally, for the spatio-temporal component, we use the definitions used above for $T_s^{(s)}$ and $T_s^{(t)}$, and write:

$$\begin{aligned} T_n^{(st)} &= \mathbf{u}_2^* \otimes \mathbf{u}_1^* \otimes \mathbf{u}_t^* \text{ and} \\ T_s^{(st)} &= [T_s^{(s)} \otimes \mathbf{u}_t^* : [\mathbf{1}_2^* \otimes \mathbf{u}_1^* : \mathbf{u}_2^* \otimes \mathbf{1}_1^* : \mathbf{u}_2^* \otimes \mathbf{u}_1^*] \otimes T_s^{(t)} : T_s^{(s)} \otimes T_s^{(t)}], \end{aligned}$$

Then, the fixed effects matrix for the spatio-temporal component is:

$$\begin{aligned} X^{(st)} &= B^{(st)} T_n^{(st)} = (B_s \otimes B_t)(\mathbf{u}_2^* \otimes \mathbf{u}_1^* \otimes \mathbf{u}_t^*) = \\ &= (B_2 \mathbf{u}_2^* \square B_1 \mathbf{u}_1^*) \otimes B_t \mathbf{u}_t^* = (\mathbf{x}_2 \square \mathbf{x}_1) \otimes \mathbf{x}_t, \end{aligned}$$

where $\mathbf{x}_k = B_k \mathbf{u}_k^*$, for $k = 1, 2, t$. And the random effects matrix is:

$$\begin{aligned} Z^{(st)} &= B^{(st)} T_s^{(st)} = \\ &= (B_s \otimes B_t)(T_s^{(s)} \otimes \mathbf{u}_t^* : [\mathbf{1}_2^* \otimes \mathbf{u}_1^* : \mathbf{u}_2^* \otimes \mathbf{1}_1^* : \mathbf{u}_2^* \otimes \mathbf{u}_1^*] \otimes T_s^{(t)} : T_s^{(s)} \otimes T_s^{(t)}) = \\ &= (B_2 U_{2s} \square B_1 U_{1n} : B_2 U_{2n} \square B_1 U_{1s} : B_2 U_{2s} \square B_1 U_{1s}) \otimes B_t \mathbf{u}_t^* : \\ &\quad (B_2 \mathbf{1}_2^* \square B_1 \mathbf{u}_1^* : B_2 \mathbf{u}_2^* \square B_1 \mathbf{1}_1^* : B_2 \mathbf{u}_2^* \square B_1 \mathbf{u}_1^*) \otimes B_t U_{ts} = \\ &\quad (B_2 U_{2s} \square B_1 U_{1n} : B_2 U_{2n} \square B_1 U_{1s} : B_2 U_{2s} \square B_1 U_{1s}) \otimes B_s U_{ts} = \\ &= (Z_2 \square X_1 : X_2 \square Z_1 : Z_2 \square Z_1) \otimes \mathbf{x}_t : [\mathbf{1}_n \square \mathbf{x}_2 : \mathbf{x}_1 \square \mathbf{1}_2 : \mathbf{x}_2 \square \mathbf{x}_1] \otimes Z_t : \\ &\quad (Z_2 \square X_1 : X_2 \square Z_1 : Z_2 \square Z_1) \otimes Z_t, \end{aligned}$$

where $\mathbf{1}_n = B_k \mathbf{u}_k^*$ for $k = 1, 2$.

Then, the complete mixed model matrices are:

$$\begin{aligned}\mathbf{X} &= [\mathbf{1}_{nt} : \mathbf{X}^{(s)} : \mathbf{X}^{(t)} : \mathbf{X}^{(st)}] \text{ and} \\ \mathbf{Z} &= [\mathbf{Z}^{(s)} : \mathbf{Z}^{(t)} : \mathbf{Z}^{(st)}],\end{aligned}$$

which are those in defined in (4.80). ■

Theorem 4.4 (Mixed model penalty for the reduced spatio-temporal S-ANOVA model). *The mixed model penalty for the reduced spatio-temporal S-ANOVA model in (4.56) is the block-diagonal matrix defined by:*

$$\mathbf{F} = \text{blockdiag}(\mathbf{F}_{(s)}, \mathbf{F}_{(t)}, \mathbf{F}_{(st)}), \quad (4.64)$$

where each block corresponds to the penalty over the smooth terms in the spatio-temporal S-ANOVA model.

Proof. As shown in (2.40), given the definition of the transformation matrix \mathbf{T} , and penalty matrix \mathbf{P} , the block-diagonal mixed model penalty \mathbf{F} is defined as $\mathbf{F} = \mathbf{T}'_s \mathbf{P} \mathbf{T}_s$. As we have demonstrated in Chapter 2, the blocks $\mathbf{F}_{(s)}$ and $\mathbf{F}_{(t)}$, are exactly the block-diagonal mixed model penalties as in a bivariate and univariate cases, with smoothing parameters λ_1 , λ_2 and λ_t , i.e.:

$$\begin{aligned}\mathbf{F}_{(s)} &= \begin{pmatrix} \lambda_2 \tilde{\Sigma}_2 \otimes \mathbf{I}_2 & & \\ & \lambda_1 \mathbf{I}_2 \otimes \tilde{\Sigma}_1 & \\ & & \lambda_1 \mathbf{I}_{c_2-2} \otimes \tilde{\Sigma}_1 + \lambda_2 \tilde{\Sigma}_2 \otimes \mathbf{I}_{c_1-2} \end{pmatrix}, \text{ and} \\ \mathbf{F}_{(t)} &= \lambda_t \tilde{\Sigma}_t.\end{aligned}$$

The last block, $\mathbf{F}_{(st)}$ is the penalty of the spatio-temporal interaction term, with smoothing parameters: τ_1 , τ_2 and τ_t . Since, in order to build this block, some columns in the bases have been removed, it requires a more detailed presentation. This block is reparameterized using that: $\mathbf{F}_{(st)} = \mathbf{T}_s^{(st)'} \mathbf{P}_{(st)} \mathbf{T}_s^{(st)}$, we obtain three sub-blocks i.e.:

$$\mathbf{F}_{(st)} = \text{blockdiag} \left(\mathbf{F}_{(st)}^{(1)}, \mathbf{F}_{(st)}^{(2)}, \mathbf{F}_{(st)}^{(3)} \right),$$

where

$$\begin{aligned}
\mathbf{F}_{(st)}^{(1)} &= \left(\mathbf{T}_s^{(s)'} \otimes \mathbf{u}_t^{*'} \right) \mathbf{P}_{(st)} \left(\mathbf{T}_s^{(s)} \otimes \mathbf{u}_t^* \right) = \\
&= \begin{bmatrix} \tau_2 \tilde{\Sigma}_2 \otimes \mathbf{I}_2 & & \\ & \tau_1 \mathbf{I}_2 \otimes \tilde{\Sigma}_1 & \\ & & \tau_2 \tilde{\Sigma}_2 \otimes \mathbf{I}_{c_1-2} + \tau_1 \mathbf{I}_{c_2-2} \otimes \tilde{\Sigma}_1 \end{bmatrix}, \\
\mathbf{F}_{(st)}^{(2)} &= \begin{pmatrix} \mathbf{1}_2^{*'} \otimes \mathbf{u}_1^{*'} \otimes \mathbf{T}_s^{(t)'} \\ \mathbf{u}_2^{*'} \otimes \mathbf{1}_1^{*'} \otimes \mathbf{T}_s^{(t)'} \\ \mathbf{u}_2^{*'} \otimes \mathbf{u}_2^{*'} \otimes \mathbf{T}_s^{(t)'} \end{pmatrix} \mathbf{P}_{(st)} \left([\mathbf{1}_2^* \otimes \mathbf{u}_1^* : \mathbf{u}_2^* \otimes \mathbf{1}_1^* : \mathbf{u}_2^* \otimes \mathbf{u}_1^*] \otimes \mathbf{T}_s^{(t)} \right) = \tau_t \mathbf{I}_3 \otimes \tilde{\Sigma}_t, \\
\mathbf{F}_{(st)}^{(3)} &= \left(\mathbf{T}_s^{(s)'} \otimes \mathbf{T}_s^{(t)'} \right) \mathbf{P}_{(st)} \left(\mathbf{T}_s^{(s)} \otimes \mathbf{T}_s^{(t)} \right) = \\
&= \tau_1 \mathbf{I}_2 \otimes \tilde{\Sigma}_1 \otimes \mathbf{I}_{c_t-2} + \tau_t \mathbf{I}_2 \otimes \mathbf{I}_{c_2-2} \otimes \tilde{\Sigma}_t + \tau_2 \tilde{\Sigma}_2 \otimes \mathbf{I}_2 \otimes \mathbf{I}_{c_3-2} + \tau_t \mathbf{I}_{c_2-2} \otimes \mathbf{I}_2 \otimes \tilde{\Sigma}_t + \\
&\quad + \tau_1 \mathbf{I}_{c_2-2} \otimes \tilde{\Sigma}_1 \otimes \mathbf{I}_{c_t-2} + \tau_2 \tilde{\Sigma}_2 \otimes \mathbf{I}_{c_1-2} \otimes \mathbf{I}_{c_3-2} + \tau_t \mathbf{I}_{c_2-2} \otimes \mathbf{I}_{c_1-2} \otimes \tilde{\Sigma}_t = \\
&= \tau_2 \tilde{\Sigma}_2 \oplus \tau_1 \tilde{\Sigma}_1 \oplus \tau_t \tilde{\Sigma}_t.
\end{aligned}$$

■

For the reduced S-ANOVA, it results of interest not only to obtain the model matrices and mixed model penalty, but also to interpret the reparameterization in terms of the recovered penalty and understand which are the constraints applied to the regression coefficients in the non-transformed (original) model. We show which are the constraints associated to the reduced S-ANOVA model in next Section.

4.4.3 Linear constraints over coefficients in the reduced spatio-temporal S-ANOVA model

For the reduced model, we can demonstrate that the reparameterization of the spatio-temporal S-ANOVA into a mixed model is equivalent to impose linear constraints over the regression coefficients: $\boldsymbol{\theta} = (\gamma, \boldsymbol{\theta}^{(s)'}, \boldsymbol{\theta}^{(t)'}, \boldsymbol{\theta}^{(st)'})'$. By 4.1, we can give an expression for the recovered penalty matrix that imposes the linear constraints in the reduced S-ANOVA model, i.e.:

$$\check{\mathbf{P}} = \mathbf{T} \boldsymbol{\Phi} \mathbf{T}' = \mathbf{K} \mathbf{P} \mathbf{K}.$$

Remark 4.6. Note that, for model in (4.56), we only have included some smooth terms (space, time and space-time interaction), i.e. main effects of latitude, longitude or interactions for latitude-time, and longitude-time were not included in the model. Hence,

for this model the constraints will be a subset of the equations shown in Table 4.1.

Proposition 4.5. *For the reduced spatio-temporal S-ANOVA model in (4.56), and given a recovered penalty $\check{P} = \mathbf{K} \mathbf{P} \mathbf{K}$. The matrix \mathbf{K} , is a constrast matrix defined as:*

$$\mathbf{K} = \begin{pmatrix} 1 & & & \\ & \mathbf{I}_{c_s} & & \\ & & \mathbf{K}_t & \\ & & & \mathbf{K}_{st} \end{pmatrix}, \quad (4.65)$$

where \mathbf{I}_{c_s} is a diagonal matrix of order $c_s = c_1 c_2$, and \mathbf{K}_t is a square matrix of order c_t and \mathbf{K}_{st} is a square matrix of order $c_s c_t$. Given the vector of coefficients $\boldsymbol{\theta} = (\gamma, \boldsymbol{\theta}^{(s)'}, \boldsymbol{\theta}^{(t)'}, \boldsymbol{\theta}^{(st)'})'$, of model (4.56), the matrix \mathbf{K} defined in (4.65), applies constraints only over the regression coefficients of the temporal and the spatio-temporal terms, (i.e. over $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\theta}^{(st)}$).

Proof. We use the relationship between \mathbf{T} and \mathbf{F} for the reduced spatio-temporal S-ANOVA model, and the penalty matrix in (4.58). Given the transformation matrix \mathbf{T} defined in 4.4, and by 4.1, we have the matrix $\mathbf{K} = \mathbf{T} \mathbf{T}'$ defined by:

$$\mathbf{K} = \mathbf{T} \mathbf{T}' = [\mathbf{T}_n : \mathbf{T}_s] \begin{bmatrix} \mathbf{T}'_n \\ \mathbf{T}'_s \end{bmatrix} = \mathbf{T}_n \mathbf{T}'_n + \mathbf{T}_s \mathbf{T}'_s.$$

Then, we have:

$$\begin{aligned} \mathbf{T}_n \mathbf{T}'_n &= \begin{bmatrix} 1 & & & \\ & \mathbf{T}_n^{(s)} \mathbf{T}_n^{(s)'} & & \\ & & \mathbf{T}_n^{(t)} \mathbf{T}_n^{(t)'} & \\ & & & \mathbf{T}_n^{(st)} \mathbf{T}_n^{(st)'} \end{bmatrix} = \\ &= \begin{bmatrix} 1 & & & \\ & \mathbf{U}_{2n} \mathbf{U}_{2n}' \otimes \mathbf{U}_{1n} \mathbf{U}_{1n}' & & \\ & & \mathbf{u}_t^* \mathbf{u}_t^{*'} & \\ & & & \mathbf{u}_2^* \mathbf{u}_2^{*'} \otimes \mathbf{u}_1^* \mathbf{u}_1^{*'} \otimes \mathbf{u}_t^* \mathbf{u}_t^{*'} \end{bmatrix}, \text{ and} \end{aligned} \quad (4.66)$$

$$\mathbf{T}_s \mathbf{T}'_s = \begin{bmatrix} 0 & & & \\ & \mathbf{T}_s^{(s)} \mathbf{T}_s^{(s)'} & & \\ & & \mathbf{T}_s^{(t)} \mathbf{T}_s^{(t)'} & \\ & & & \mathbf{T}_s^{(st)} \mathbf{T}_s^{(st)'} \end{bmatrix}, \quad (4.67)$$

where in (4.67):

$$\begin{aligned}
\mathbf{T}_s^{(s)} \mathbf{T}_s^{(s)'} &= \mathbf{U}_{2s} \mathbf{U}_{2s}' \otimes \mathbf{U}_{1n} \mathbf{U}_{1n}' + \mathbf{U}_{2n} \mathbf{U}_{2n}' \otimes \mathbf{U}_{1s} \mathbf{U}_{1s}' + \mathbf{U}_{2s} \mathbf{U}_{2s}' \otimes \mathbf{U}_{1s} \mathbf{U}_{1s}', \\
\mathbf{T}_s^{(t)} \mathbf{T}_s^{(t)'} &= \mathbf{U}_{ts} \mathbf{U}_{ts}', \\
\mathbf{T}_s^{(st)} \mathbf{T}_s^{(st)'} &= \mathbf{T}_s^{(s)} \mathbf{T}_s^{(s)'} \otimes \mathbf{u}_t^* \mathbf{u}_t^{*'} + [\mathbf{1}_2^* \mathbf{1}_2^{*'} \otimes \mathbf{u}_1^* \mathbf{u}_1^{*'} + \mathbf{u}_2^* \mathbf{u}_2^{*'} \otimes \mathbf{1}_1^* \mathbf{1}_1^{*'} + \mathbf{u}_2^* \mathbf{u}_2^{*'} \otimes \mathbf{u}_1^* \mathbf{u}_1^{*'}] \otimes \mathbf{T}_s^{(t)} \mathbf{T}_s^{(t)'} + \\
&\quad + \mathbf{T}_s^{(s)} \mathbf{T}_s^{(s)'} \otimes \mathbf{T}_s^{(t)} \mathbf{T}_s^{(t)'}.
\end{aligned}$$

Given (4.66) and (4.67), we rewrite \mathbf{K} as:

$$\mathbf{K} = \begin{bmatrix} 1 & & \\ & \mathbf{I}_{c_2} \otimes \mathbf{I}_{c_1} & \\ & \underbrace{\mathbf{u}_t^* \mathbf{u}_t^{*'} + \mathbf{U}_{ts} \mathbf{U}_{ts}'}_{(a)} & \\ & & \underbrace{\mathbf{T}_n^{(st)} \mathbf{T}_n^{(st)'} + \mathbf{T}_s^{(st)} \mathbf{T}_s^{(st)'}}_{(b)} \end{bmatrix}. \quad (4.68)$$

As we have shown in 4.3, the expression (a) in (4.68) is a centering matrix, such that, it can be replaced by $\mathbf{K}_t = (\mathbf{I}_{c_t} - \mathbf{1}\mathbf{1}'/c_t)$. Note that, in the reduced S-ANOVA case, the constrast matrix has not a direct interpretation as in the full ANOVA model. Given that, the term (b) in (4.68) has not a interpretable expression by itself. Using the results in Proof of 4.3, we can simplify \mathbf{K}_{st} as:

$$\begin{aligned}
\mathbf{K}_{st} &= [\mathbf{u}_2^* \mathbf{u}_2^{*'} \otimes \mathbf{u}_1^* \mathbf{u}_1^{*'} + \mathbf{I}_{c_2} \otimes \mathbf{U}_{1s} \mathbf{U}_{1s}' + \mathbf{U}_{2s} \mathbf{U}_{2s}' \otimes \mathbf{U}_{1n} \mathbf{U}_{1n}'] \otimes \mathbf{u}_t^* \mathbf{u}_t^{*'} + \\
&\quad + [(\mathbf{u}_2^* \mathbf{u}_2^{*'} + \mathbf{U}_{2s} \mathbf{U}_{2s}') \otimes \mathbf{U}_{1n} \mathbf{U}_{1n}'] \otimes \mathbf{U}_{ts} \mathbf{U}_{ts}' + \\
&\quad + [\mathbf{1}_2^* \mathbf{1}_2^{*'} \otimes \mathbf{u}_1^* \mathbf{u}_1^{*'} + \mathbf{I}_{c_2} \otimes \mathbf{U}_{1s} \mathbf{U}_{1s}'] \otimes \mathbf{U}_{ts} \mathbf{U}_{ts}',
\end{aligned} \quad (4.69)$$

that is a square matrix of order $c_s c_t$. ■

In order to obtain the constraints over the regression coefficients $\boldsymbol{\theta}$ applied by the constrast matrix \mathbf{K} defined in (4.65), we will construct the recovered penalty matrix $\check{\mathbf{P}}$, for the reduced spatio-temporal S-ANOVA.

Theorem 4.5 (Recovered penalty matrix for the reduced spatio-temporal S-ANOVA model). *The recovered penalty for the reduced spatio-temporal S-ANOVA model in (4.56) is a block-diagonal, given by:*

$$\check{\mathbf{P}} = \mathbf{K} \mathbf{P} \mathbf{K} = \text{blockdiag}(0, \check{\mathbf{P}}_{(s)}, \check{\mathbf{P}}_{(t)}, \check{\mathbf{P}}_{(st)}), \quad (4.70)$$

Proof. Given matrix \mathbf{K} in (4.68) and penalty \mathbf{P} in (4.58), we have that for each block of

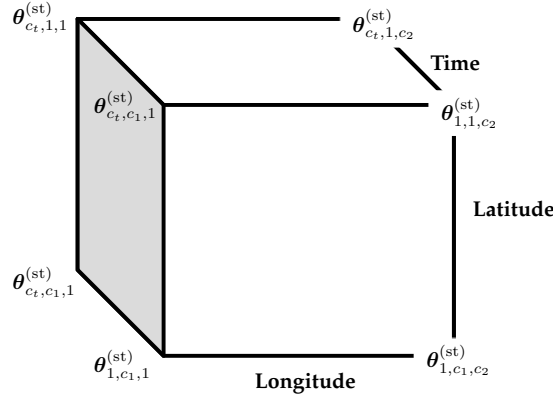


Figure 4.5: Array $\Theta^{(st)}$ of coefficients for the space-time interaction, of dimensions $c_t \times c_1 \times c_2$.

(4.70), we obtain:

$$\check{P}_{(s)} = I_{c_s} P_{(s)} I_{c_1} = P_{(s)}, \quad (4.71)$$

$$\check{P}_{(t)} = K_t P_{(t)} K_t, \text{ and} \quad (4.72)$$

$$\check{P}_{(st)} = K_{st} P_{(st)} K_{st}, \quad (4.73)$$

Note that, (4.71) is the original penalty for the spatial coefficients, and therefore, the vector of coefficients for the spatial part is not constrained. The recovered penalty for the temporal coefficient $\check{P}_{(t)}$ in (4.72), centers the regression coefficients for the temporal component, i.e. $\theta^{(t)}$. Finally, the recovered penalty matrix in (4.73), corresponds to the spatio-temporal coefficients. For given K_{st} , and penalty matrix over $P_{(st)}$, we have:

$$\begin{aligned} \check{P}_{(st)} &= K_{st} P_{(st)} K_{st} = \\ &= \left([u_2^* u_2^{*'} \otimes u_1^* u_1^{*'} + I_{c_2} \otimes U_{1s} U_{1s}' + U_{2s} U_{2s}' \otimes U_{1n} U_{1n}'] \otimes u_t^{*'} u_t^* + \right. \\ &\quad \left. + [(u_2^* u_2^{*'} + U_{2s} U_{2s}') \otimes U_{1n} U_{1n}'] \otimes U_{ts} U_{ts}' + \right. \\ &\quad \left. + [1_2^* 1_2^{*'} \otimes u_1^* u_1^{*'} + I_{c_2} \otimes U_{1s} U_{1s}'] \otimes U_{ts} U_{ts}' \right) \left. \right\} K_{(st)} \\ &\quad \left(\tau_2 D_2' D_2 \otimes I_{c_1} \otimes I_{c_t} + \tau_1 I_{c_2} \otimes D_1' D_1 \otimes I_{c_t} + \tau_3 I_{c_2} \otimes I_{c_1} \otimes D_t' D_t \right) \left. \right\} P_{(st)} \\ &\quad \left([u_2^* u_2^{*'} \otimes u_1^* u_1^{*'} + I_{c_2} \otimes U_{1s} U_{1s}' + U_{2s} U_{2s}' \otimes U_{1n} U_{1n}'] \otimes u_t^{*'} u_t^* + \right. \\ &\quad \left. + [(u_2^* u_2^{*'} + U_{2s} U_{2s}') \otimes U_{1n} U_{1n}'] \otimes U_{ts} U_{ts}' + \right. \\ &\quad \left. + [1_2^* 1_2^{*'} \otimes u_1^* u_1^{*'} + I_{c_2} \otimes U_{1s} U_{1s}'] \otimes U_{ts} U_{ts}' \right) \left. \right\} K_{(st)} \quad (4.74) \end{aligned}$$

Using the next identities (for $k = 1, 2, t$):

$$U_{kn} U_{kn}' D_k' D_k U_{kn} U_{kn}' = u_k^* u_k^{*'} D_k' D_k u_k^* u_k^{*'} = 0_{c_k}, \text{ and also}$$

$$U_{ks} U_{ks}' D_k' D_k U_{ks} U_{ks}' = D_k' D_k,$$

we simplify the penalty matrix in (4.74) as:

$$\begin{aligned} \check{P}_{(st)} = & \tau_1 \underbrace{I_{c_2} \otimes D_1' D_1 \otimes K_t}_{(a)} + \tau_2 \underbrace{D_2' D_2 \otimes I_{c_1} \otimes K_t}_{(b)} + \\ & + \tau_t \underbrace{(I_{c_2} \otimes I_{c_1} - \mathbf{1}\mathbf{1}_2'/c_2 \otimes \mathbf{1}\mathbf{1}_1'/c_1) \otimes D_t' D_t}_{(c)}, \end{aligned} \quad (4.75)$$

where (a) and (b) impose constraints over each dimension of the array $\Theta^{(st)}$ (See Figure 4.5). The last term (c), can be rewritten as $K_s \otimes D_t' D_t$, where K_s is a centering matrix over dimension $c_s = c_1 c_2$, i.e.: $K_s = (I_{c_s} - \mathbf{1}\mathbf{1}_s'/c_s)$. ■

Theorem 4.6 (Linear constraints in the reduced spatio-temporal S-ANOVA model). *The linear constraints applied over the regression coefficients vector θ , in the reduced spatio-temporal S-ANOVA model in (4.56) are:*

$$\sum_{t=1}^{c_t} \theta_t^{(t)} = 0, \quad \text{and} \quad (4.76)$$

$$\sum_i^{c_1} \theta_{t,ij}^{(st)} = \sum_j^{c_2} \theta_{t,ij}^{(st)} = \sum_i^{c_1} \sum_j^{c_2} \theta_{t,ij}^{(st)} = 0. \quad (4.77)$$

Proof. First, the linear constraint (4.76), is applied on the temporal vector of coefficients $\theta^{(t)}$, and it is obtained from the recovered penalty matrix $\check{P}_{(t)}$ defined in (4.72) and the matrix K_t obtained in (4.65), i.e., the penalty over $\theta^{(t)}$ is:

$$\theta^{(t)'} \check{P}_{(t)} \theta^{(t)} = \theta^{(t)'} (K_t P_{(t)} K_t) \theta^{(t)} = \check{\theta}^{(t)'} P_{(t)} \check{\theta}^{(t)}, \quad (4.78)$$

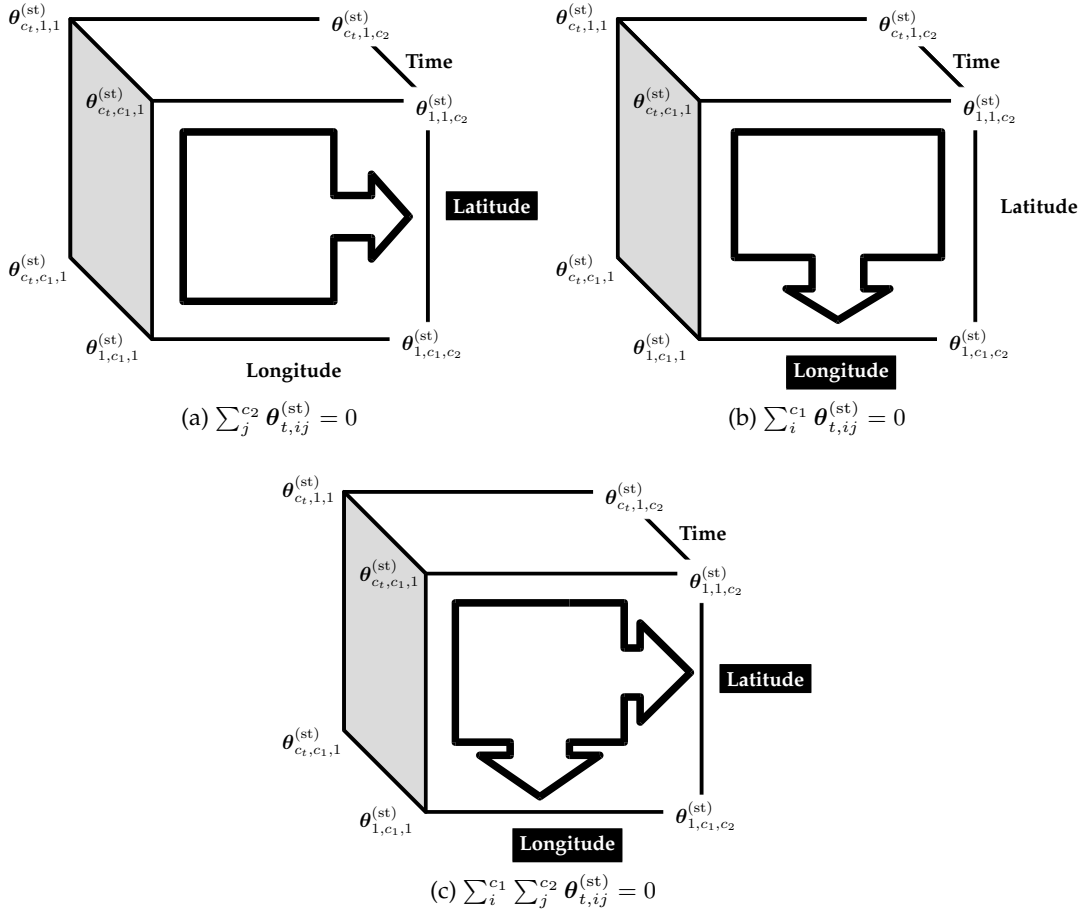
where $\check{\theta}^{(t)} = K_t \theta^{(t)}$ is the centered vector of regression coefficients for the temporal smooth term. Given that, (4.78) might be written as:

$$\sum_{t=1}^{c_t} \theta_t^{(t)'} K_t D_t' D_t K_t \theta_t^{(t)},$$

that it is equivalent to impose the linear constraint in (4.76). Secondly, the set of linear constraints in (4.77) are obtained as follows: given the recovered penalty matrix $\check{P}_{(st)}$ defined in (4.75), we have that, each part of $\check{P}_{(st)}$, corresponds to applying penalties over the P -spline regression array of coefficients, $\Theta^{(st)}$, i.e.:

- The term (a) in (4.75) is equivalent to:

$$\sum_{j=1}^{c_2} \theta_{t,ij}^{(st)'} K_t D_1' D_1 K_t \theta_{t,ij}^{(st)},$$

Figure 4.6: Restrictions over the array $\Theta^{(st)}$, in spatial dimensions

and it implies the linear restriction $\sum_{j=1}^{c_2} \theta_{t,ij}^{(st)} = 0$.

- The term (b) in (4.75) is equivalent to:

$$\sum_{i=1}^{c_1} \theta_{t,ij}^{(st)'} K_t D_2' D_2 K_t \theta_{t,ij}^{(st)},$$

and it implies the linear restriction $\sum_{i=1}^{c_1} \theta_{t,ij}^{(st)} = 0$.

- And the term (c) in (4.75) is equivalent to:

$$\sum_{i=1}^{c_1} \sum_{j=1}^{c_2} \theta_{t,ij}^{(st)'} K_s D_t' D_t K_s \theta_{t,ij}^{(st)},$$

and it implies the linear restriction $\sum_i^{c_1} \sum_j^{c_2} \theta_{t,ij}^{(st)} = 0$.

■

Figure 4.6 illustrates the linear constraints showed in (4.77) in terms of the imposed constraints over the dimensions of the array of coefficients $\Theta^{(st)}$.

4.4.4 Analysis of air pollution levels in Europe

A repeated exposure to ozone pollution at ground-level may cause important damages to human health (including asthma, reduced lung capacity or susceptibility to respiratory illnesses), ecosystems and agricultural crops. The formation of ozone is increased by hot weather and in urban industrial areas, and the concentrations over Europe also present a wide variation and large differences due to climate conditions over the continent. Therefore, it is expected that ozone concentrations around Europe present a spatio-temporal pattern.

The harmful effects of ozone have become an important issue the development of new policies. The European Environment Agency (EEA) has established a program to monitor changes in ozone levels in the last decade. The EEA presents annual evaluation reports of ground-level ozone pollution in Europe from April-September, based on information submitted to the European Commission on ozone in ambient air. According to this annual reports, although emissions of ozone precursors have been reduced over the last decade, ozone pollution levels has not changed significantly in the period 1999-2005. The analysis of the data will confirm this statement, but it will show that different countries reach the largest values of ozone at different time points.

We analyzed monthly averages of air pollution by ground-level ozone (in $\mu g/m^3$ units) over Europe from January 1999 to December 2005. The data were collected in 43 monitoring stations in 15 EU countries. Following the methodology described in previous sections, we fitted 3 models to the data: (i) *spatio-temporal S-ANOVA model*; (ii) *3d interaction model* and (iii) *space-time additive model*. The three models formulation are then:

- i. **S-ANOVA:** $f_s(\mathbf{x}_1, \mathbf{x}_2) + f_t(\mathbf{x}_t) + f_{s,t}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_t)$,
- ii. **Interaction:** $f_{s,t}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_t)$, and
- iii. **Additive:** $f_s(\mathbf{x}_1, \mathbf{x}_2) + f_t(\mathbf{x}_t)$

In order to fit the models, we set up the B -splines bases, using the following parameters: (1) the number of (equidistant) internal knots, ndx ; (2) the degree of the P -spline, $bdeg$; and the order of the penalty, $pord$. We selected one knot for every four or five observations. The parameters were: $bdeg = 3$ (cubic B -splines), $pord = 2$ (second order penalty) and $ndx_{(s)} = (10, 10)$ for both spatial dimensions, and $ndx_{(t)} = 21$ for time, in order to have enough flexibility to capture the seasonal time trend. Then, the spatial bases B_1 and B_2 are of dimension 43×13 , and B_t has dimension 84×24 .

Table 4.2: AIC and estimated degrees of freedom of fitted models.

Model	AIC	Dev	ED
S-ANOVA	14280.73	13548.67	366.03
Interaction	14537.22	13007.12	765.05
Additive	16506.28	16374.32	65.98

The mixed model formulation is straightforward following the methodology proposed in the paper: we construct matrices \mathbf{X} and \mathbf{Z} , and the block-diagonal penalty \mathbf{F} for each model. We compared the performance of the models in terms of the Akaike Information Criteria (**AIC**), calculated as

$$\text{AIC} = \text{Dev} + 2 \text{ED},$$

where Dev is the deviance calculated as $\text{Dev} = \sum_i^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2$, and ED are the effective degrees of freedom of the model, measured as the trace of the hat-matrix, as shown in Section 2.1.2.

The results are summarized in Table 4.2. There is a superior performance of S-ANOVA and interaction models with respect to the additive model. This could be expected since it is unrealistic to force the spatial pattern of ozone concentrations to increase and decrease similarly in all locations. The interaction model, although giving a better fit, uses a large amount of effective degrees of freedom. This is due to the fact that model has a single smoothing parameter for the temporal component. Then, the strong seasonal trend forces the model to use a small smoothing parameter (large ED). The S-ANOVA model performs better. It uses less degrees of freedom because the model allows a different the amount of smoothing in the additive temporal term and the spatio-temporal component, and, as we could expect, the temporal smoothing in the interaction does not need to be so strong. These results in a more parsimonious model.

Figure 4.7a shows the smoothed spatial surface for the ozone levels of the S-ANOVA model. The estimated spatial trend surface reflects a non-uniform picture across Europe, since the highest concentrations are observed in Southern Europe in Mediterranean countries as Spain, France and Italy, and the lowest levels are in North West Europe and the UK. The seasonal cycle of ozone levels is captured by the temporal trend shown in Figure 4.7b, where the highest levels are recorded during spring and summer months (April-September). The highest peak corresponds to the heat-wave occurred in Europe during summer 2003. The spatio-temporal S-ANOVA model also allows the explicit modelling of the space-time interaction in addition to the spatial and temporal trends. Figure 4.8 shows this interaction from March to August 2002. As it can be seen from

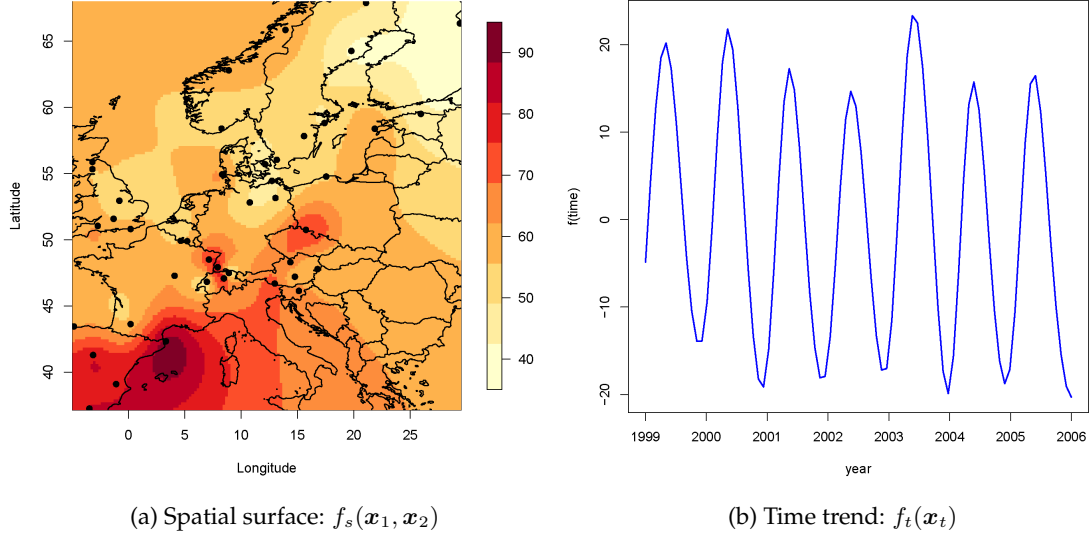


Figure 4.7: Spatial and temporal smooth terms for S-ANOVA model.

the sequence of figures, there are differences between north west and southern and Mediterranean countries throughout the summer period.

The differences between additive and S-ANOVA models can be seen in Figure 4.9. We plotted the fitted values for four different monitoring stations against the raw ozone levels data. The additive model, ignores the interaction and assumes a spatial smooth surface over all monitoring stations that remains constant over time. The fitted values vary smoothly according to a seasonal pattern, but maintain the same differences among locations (Figure 4.9a). In contrast, the spatio-temporal S-ANOVA model fit, is able to capture the individual characteristics of the stations throughout time. Figure 4.9b shows the particular phase and amplitude given the geographic and seasonal inter-annual variations of four monitoring stations. The high and low season for ozone concentrations are different, depending on the location, and the cycle changes over time.

4.5 Smooth-ANOVA models and nested B -spline bases

The S-ANOVA methodology developed in previous Section, may be constrained by the number of parameters to be estimated. In some cases, we are more interested in providing more flexibility for the main effects rather than for the interactions. Let us consider a full S-ANOVA model with two covariates, i.e. $f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + f_{1,2}(\mathbf{x}_1, \mathbf{x}_2)$, with regression basis \mathbf{B}_k , of dimension $n_k \times c_k$, for $k = 1, 2$, where c_k is the number of columns

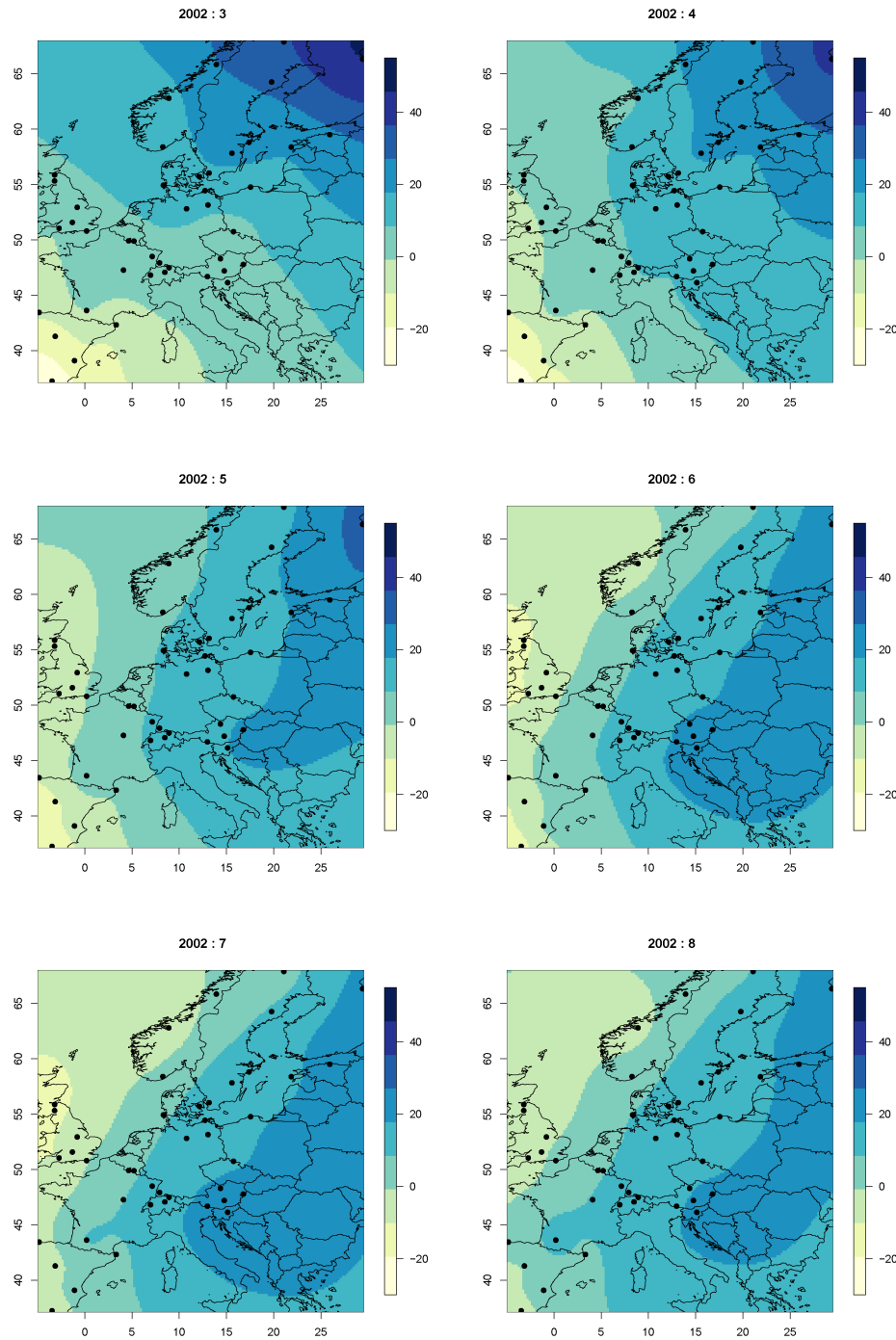


Figure 4.8: Spatio-temporal interaction fit for the spatio-temporal S-ANOVA model, from March to August 2002.

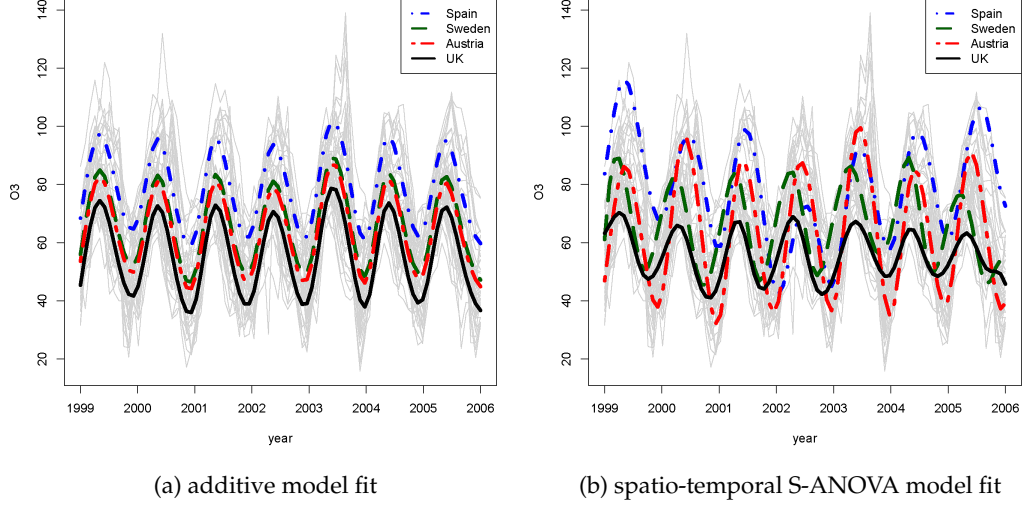


Figure 4.9: Comparison of fitted values for monitoring stations in Spain, Sweden, Austria and UK.

of B_k . The natural choice in the S-ANOVA modelling, is to consider the same marginal bases for the additive terms ($f_1(x_1) + f_2(x_2)$), and for the interaction ($f_{1,2}(x_1, x_2)$), i.e. B_1 and B_2 as shown in (4.19), to ensure that both additive and S-ANOVA models are strictly nested (see Wood, 2006a, Chapter 4). However, in some cases, the size of the interaction basis $B_2 \otimes B_1$, is very large and the number of parameters to estimate for the interaction are $c_1 c_2$. Then the total number of parameters to estimate are: “constant + $c_1 + c_2 + c_1 c_2$ ”, that may lead to computational limitations. A simple solution, is to reduce the number of parameters for the interaction terms. This idea can be explained by analogy to classical ANOVA models, where in general the main effects are more significant than interactions.

A simple solution is to reduce the size of the basis functions for the interaction. For model (4.18), we can replace from the regression basis in (4.19), the last block for the interaction by:

$$\tilde{B}_2 \otimes \tilde{B}_1,$$

where \tilde{B}_1 and \tilde{B}_2 , are lower rank basis functions of dimensions $n_1 \times \tilde{c}_1$ and $n_2 \times \tilde{c}_2$, respectively, i.e.:

$$\text{rank}(\tilde{B}_k) < \text{rank}(B_k) \rightarrow \tilde{c}_k < c_k, \text{ for } k = 1, 2,$$

Then, reducing the rank of the B -spline basis for the interaction, we reduce the number

of parameters to estimate for the interaction to $\tilde{c}_1\tilde{c}_2 < c_1c_2$.

However, taking a reduced basis of arbitrary size will yield a model that will not be nested on the additive model (i.e. $f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2)$), and so, the comparison between additive and interaction models will not be straightforward.

We propose the use of *nested B-spline bases* for the interaction term, i.e., basis such that the space spanned by $\tilde{\mathbf{B}}_k$, is a subset of the space spanned by \mathbf{B}_k , and so, the hierarchical nature of the models is preserved. Using these nested B-spline bases, the identifiability constraints shown in Section 4.2 remain the same, and the total number of parameters is greatly reduced.

The way to ensure that the new basis is nested on the original is to use a number of knots that is a divisor of the number of knots used in the original basis, i.e.:

$$\#\text{knots}(\tilde{\mathbf{B}}_k) = \frac{\#\text{knots}(\mathbf{B}_k)}{\text{div}_k} \Rightarrow \text{span}(\tilde{\mathbf{B}}_k) \subset \text{span}(\mathbf{B}_k),$$

and div_k is any divisor of the number of knots used to construct \mathbf{B}_k (Figure 4.10 shows an example of basis with 8 and 4 knots, with a divisor of $\text{div} = 2$). In next Section, we illustrate the idea of nested B-spline bases for spatio-temporal data.

4.5.1 Nested B-spline basis for spatio-temporal data

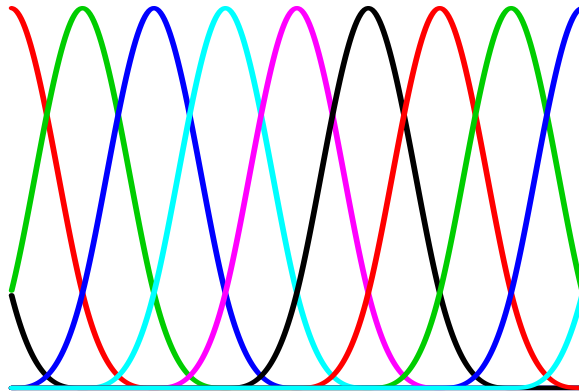
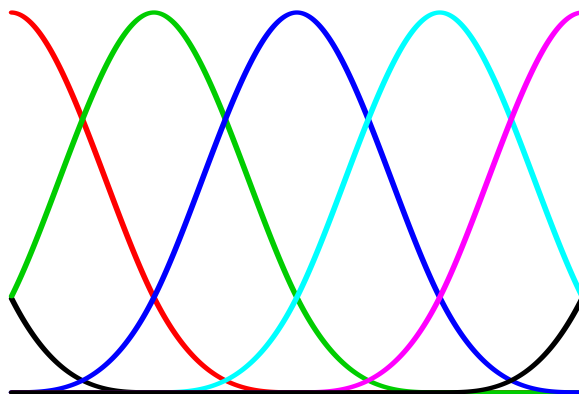
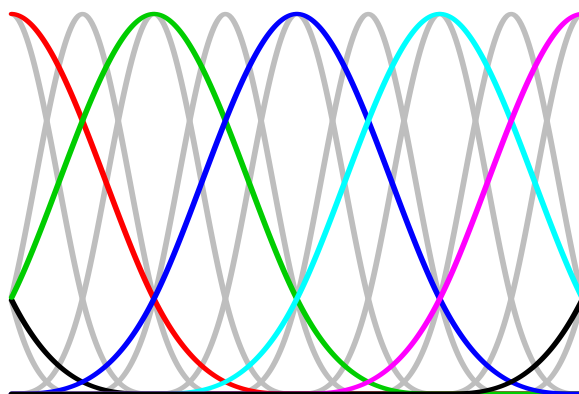
The use of the nested B-spline bases results very attractive in the spatio-temporal data modelling. Environmental data often presents a strong seasonal trend, and the use of the reduced spatio-temporal S-ANOVA model in (4.50), may require that the basis \mathbf{B}_t in (4.57) has to be large (between 20 and 40 equidistant knots) in order to have enough degrees of freedom to capture the temporal structure. As a consequence, the number of parameters in the interaction (those associated with the Tensor product $\mathbf{B}_s \otimes \mathbf{B}_t$), could easily be of the order of thousands, and the computational burden prohibitive.

We propose to replace the regression basis for the reduced spatio-temporal S-ANOVA in (4.57), by:

$$\mathbf{B} = [\mathbf{1}_{nt} : \mathbf{B}_s \otimes \mathbf{1}_t : \mathbf{1}_n \otimes \mathbf{B}_t : \mathbf{B}_s \otimes \tilde{\mathbf{B}}_t]. \quad (4.79)$$

The only difference is in the last block, where $\tilde{\mathbf{B}}_t$ is a nested B-spline basis with the temporal main effect \mathbf{B}_t . The use of the nested B-spline basis, leaves unchanged the mixed model reparameterization showed in Section 4.4 for the reduced S-ANOVA model. Then, the penalty matrix is defined as in (4.58), with spatio-temporal penalty block:

$$\tilde{\mathbf{P}}_{(st)} = \tau_2 \mathbf{D}'_2 \mathbf{D}_2 \otimes \mathbf{I}_{c_1} \otimes \mathbf{I}_{\tilde{c}_t} + \tau_1 \mathbf{I}_{c_2} \otimes \mathbf{D}'_1 \mathbf{D}_1 \otimes \mathbf{I}_{\tilde{c}_t} + \tau_3 \mathbf{I}_{c_2} \otimes \mathbf{I}_{c_1} \otimes \tilde{\mathbf{D}}'_t \tilde{\mathbf{D}}_t.$$

(a) *B*-spline basis constructed with 8 knots(b) *B*-spline basis constructed with 4 knots(c) overlapping both *B*-spline basesFigure 4.10: Illustrative example of two nested *B*-spline bases, with $d = 2$.

where $\tilde{\mathbf{D}}_t' \tilde{\mathbf{D}}_t$, is a penalty matrix for the nested B -spline basis coefficients, of order $\tilde{c}_t \times \tilde{c}_t$. The SVD over the penalty matrix $\tilde{\mathbf{D}}_t' \tilde{\mathbf{D}}_t$ is:

$$\tilde{\mathbf{D}}_t' \tilde{\mathbf{D}}_t = \tilde{\mathbf{U}}_t \mathbf{\Psi}_t \tilde{\mathbf{U}}_t',$$

where $\tilde{\mathbf{U}}_t$ is the matrix of eigenvectors, and $\mathbf{\Psi}_t$ is a diagonal matrix of eigenvalues $\mathbf{\Psi}_t = \text{blockdiag}(\mathbf{0}_{q_t}, \tilde{\mathbf{\Psi}}_t)$, where $\tilde{\mathbf{\Psi}}_t$ are the positive eigenvalues. Given these definitions, the mixed model bases are similar to those obtained in the reduced S-ANOVA model in (4.80), but the last block for the space-time interaction, includes the random effects matrix $\tilde{\mathbf{Z}}_t$, defined as $\tilde{\mathbf{Z}}_t = \tilde{\mathbf{B}}_t \tilde{\mathbf{U}}_{ts}$, where $\tilde{\mathbf{U}}_{ts}$ is the sub-matrix of eigenvectors corresponding to the positive eigenvalues of the SVD over $\tilde{\mathbf{D}}_t' \tilde{\mathbf{D}}_t$. Finally, the new mixed model bases are:

$$\begin{aligned} \mathbf{X} &= [\mathbf{X}_s \otimes \mathbf{1}_t : \mathbf{1}_n \otimes \mathbf{x}_t : \mathbf{x}_s \otimes \mathbf{x}_t]. \\ \mathbf{Z} &= [\mathbf{Z}_s \otimes \mathbf{1}_t : \mathbf{1}_n \otimes \mathbf{Z}_t : \mathbf{Z}_s \otimes \mathbf{x}_t : (\mathbf{x}_1 : \mathbf{x}_2 : \mathbf{x}_s) \otimes \tilde{\mathbf{Z}}_t : \mathbf{Z}_s \otimes \tilde{\mathbf{Z}}_t]. \end{aligned}$$

Following the same procedure in Section 4.4, we obtain the block-diagonal mixed model penalty matrix as shown in (4.64), with diagonal matrix of eigenvectors $\tilde{\mathbf{\Psi}}_t$, for the space-time interaction. Note that, the use of the nested B -spline basis, only affects to the size of the basis and the number of coefficients, and therefore the linear constraints are applied as shown in Section 4.4.3. We illustrate an application of the S-ANOVA methodology with nested B -spline basis in the spatio-temporal smoothing context in next Section.

US temperature data

We apply the reduced spatio-temporal S-ANOVA model with nested B -spline basis, to the analysis of monthly average temperatures (in $^{\circ}\text{F}$) across the U.S. between January 1995 and December 2004. Figure 4.11a shows the spatial locations (a total number of 136 U.S. cities), and Figure 4.11b the time series data. For the spatial smooth term, $f_s(\mathbf{x}_1, \mathbf{x}_2)$, we constructed the B -spline bases (\mathbf{B}_1 and \mathbf{B}_2) with 10 equidistant knots for each x_1 and x_2 coordinates, to cover the spatial domain. For the time trend, we start by constructing \mathbf{B}_t , with 4 knots per year, with a total of 30 knots across x_t (otherwise the seasonal effect would not be captured). Fitting the reduced S-ANOVA model in (4.50) (i.e. the non-nested model), without the nested basis, led to a total of 5779 parameters, and the size of the matrices involved made the fit of the model computationally very intensive in standard requirements PCs.

In order to reduce the computational burden, we decreased the number of knots in the time basis \mathbf{B}_t in model (4.50). However, this leads us to oversmoothing the seasonal

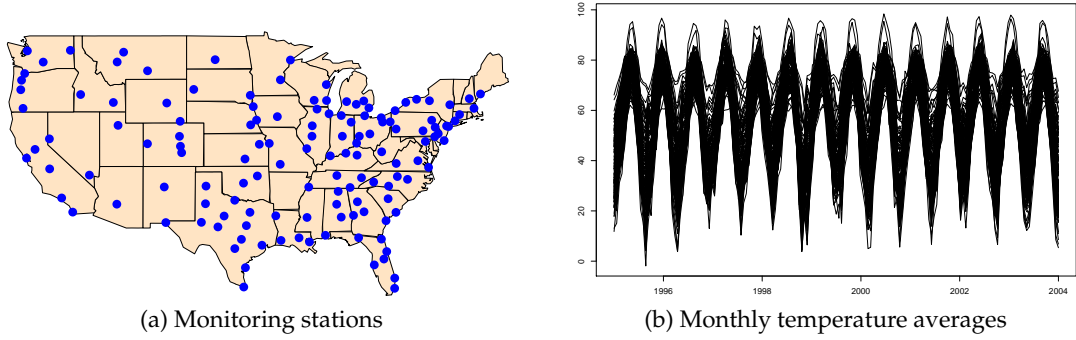


Figure 4.11: U.S. monthly average temperature (in $^{\circ}\text{F}$) data of 136 cities from January 1995 and December 2004 ($t = 120$ time points). The total number of observations is 16320.

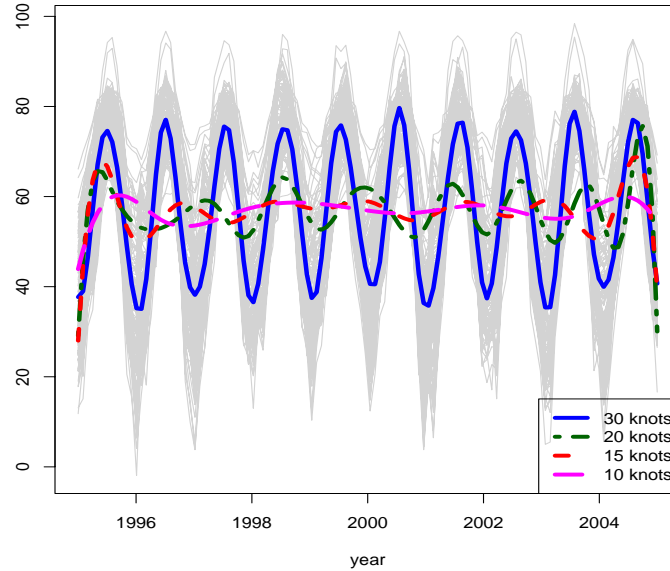


Figure 4.12: U.S. temperature data time trend: $f_t(x_t)$. Fitted with 30,20,15 and 10 knots in the construction of the B -spline basis B_t , in the reduced S-ANOVA without nested basis in the space-time interaction.

trend. Figure 4.12, shows the smoothed time trends estimated using the reduced spatio-temporal model (4.50), with different number of knots in the construction of B_t . The Figure illustrates the need of choosing a large marginal basis for temporal trend. Using less than 30 knots in the temporal basis B_t , the temporal trend, $f_t(x_t)$, is not flexible enough to capture the seasonality. Figure 4.13 shows the estimated smooth spatial trend with the reduced S-ANOVA model.

For the nested bases approach, we construct the nested B -spline time basis \tilde{B}_t ,

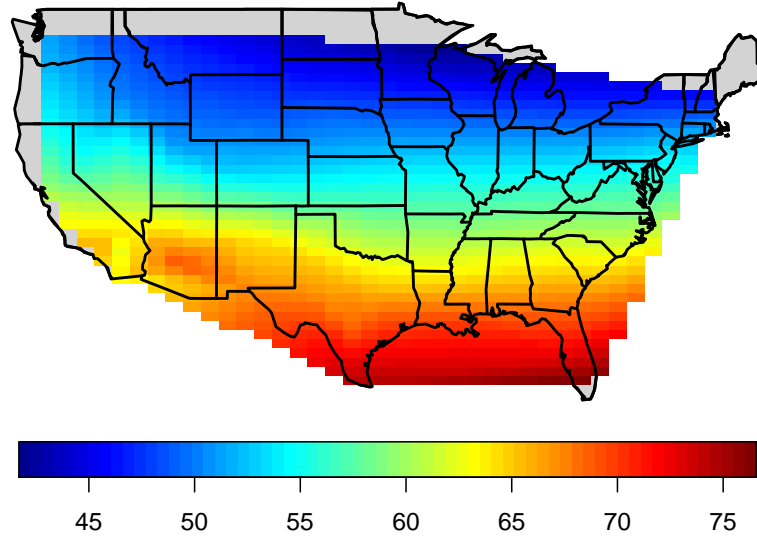


Figure 4.13: U.S. temperature data spatial effect: $f_s(x_1, x_2)$. The Figure shows the a south to north spatial pattern.

with 15, 10 and 6 knots (i.e. with $div_t = \{2, 3, 5\}$). Table 4.3, shows the % of reduction achieved by using the nested basis in terms of total number of parameters estimated. For the CPU time comparative study, and given that the estimation of the smoothing parameters and σ are done by maximization of the REML function, and it may depend on the starting values, we considered the estimation of the models using REML for fixed smoothing parameters and σ , and 20 iterations of the algorithm (which takes around 2 hours).

Table 4.3: Summary of nested models with reduced number of knots, respect to non-nested model with 30 knots in B_t and 5779 parameters.

# param.	# reduced knots	reduction (%)	CPU time reduction (%)
3075	15	47%	82%
2399	10	58%	94%
1723	6	70%	97%

We compared the performance of the models in terms of the Akaike Information Criteria (AIC), and the model degrees of freedom. The results obtained are summarized in Table 4.4, although the AIC increases with the reduction of the number of parameters, this reduction supposed a decrease of a 8% and did not significantly affect the goodness-of-fit of the model. Therefore, the selection of a more parsimonious nested model was a reasonable choice. Further research in terms of goodness-of-fit and sensi-

tivity analysis is needed in this approach.

Table 4.4: Comparison of AIC and df of non-nested and nested basis models.

Model basis	AIC	edf
non-nested	45424.37	878.18
nested		
15 knots	49374.31	158.35
10 knots	49565.24	147.84
6 knots	49648.49	109.40

4.6 Further considerations

In both of the examples presented in this Section, we have considered environmental data (ozone levels and temperature data), where the seasonal effect is modelled by a smooth term $f_t(\mathbf{x}_t)$. In this type of studies, the strong periodic effects have an important impact on the environmental response. In environmental time series analysis, it is common to consider that most of the periodicity and its source is not stochastic, and instead can be assumed to be deterministic. Then, following a time series approach, the seasonal effect can be modelled parametrically and removed before other effects are estimated. These models are known as *harmonic regression models*. Formally, consider $\mathbf{y}_t^{(i)}$ as the response at the i^{th} monitoring station (i.e. a single time series), a simple harmonic regression model can be written as:

$$\mathbf{y}_t^{(i)} = \beta_0 + \gamma \sin\left(\frac{2\pi(\mathbf{x}_t - \varphi)}{p}\right) + \epsilon_t, \quad \text{for } t = 1, \dots, T \quad (4.80)$$

where \mathbf{x}_t is the temporal covariate (in days, months, or years), φ is the *phase angle* in radians, such that $-1 \leq \sin(\varphi) \leq 1$, and p is the *period* (the time of one cycle), and ϵ_t is the error term, that in the case of no temporal correlation is *i.i.d.*, $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. [Figure 4.14](#) summarizes the components parameters of the sinusoidal regression model in (4.80). A simple model, would be to incorporate this model as a semi-parametric model:

$$\mathbf{y}_t^{(i)} = \underbrace{\beta_0 + \gamma \sin(2\pi(\mathbf{x}_t - \varphi)/p)}_{\text{harmonic model}} + \underbrace{\mathbf{X}\beta + \mathbf{Z}\alpha}_{P\text{-spline}} + \epsilon_t.$$

From the *P-spline* approach, [Eilers et al. \(2008\)](#) and [Marx et al. \(2010\)](#), considered the incorporation of specific smooth structures to capture the seasonal/cyclic patterns in periodic data. They use harmonic smooth terms as *varying-coefficient models* ([Hastie and](#)

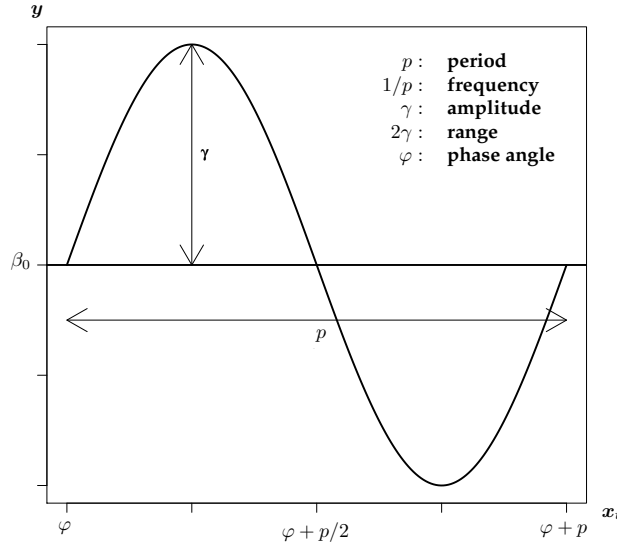


Figure 4.14: Sine function: $\beta_0 + \gamma \sin(2\pi(x_t - \varphi)/p)$.

Tibshirani, 1993) for P -GLM Poisson regression of seasonal counts. For the univariate case, we can model the response variable as:

$$\mathbf{y}_t^{(i)} = v_t + f_t \cos(\omega t) + g_t \sin(\omega t) + \epsilon_t, \quad (4.81)$$

where $\omega = 2\pi/p$, for period p . The smooth trend is modelled by v_t , and f_t and g_t are smooth terms for the amplitudes of the cosine and sine waves. Model in (4.81) is then expressed as a P -spline (additive) varying-coefficient model (see Eilers and Marx, 2002). Eilers et al. (2008) and Marx et al. (2010) extend the methodology to bivariate smoothing of age-time incidence tables using GLAM methods.

For periodic data, Eilers and Marx (2004) proposed the use of specific bases to consider the periodicity of the data. This *circular* or *harmonic* B -spline basis are such that in the linear axis both ends match at the boundaries. These bases can be constructed wrapping at the first and last knot locations (see also Wood, 2006a, Ch. 4). Figure 4.15 compares standard and cyclic B -spline bases. The difference penalty D is then also changed wrapping it around in the same way as for the B -spline basis.

Eilers and Marx (2004) also consider the use of *specialized* or *designer* penalties, changing the usual difference penalty on adjacent coefficients, i.e. $\lambda \sum (\theta_j - 2\theta_{j-1} + \theta_{j-2})^2$, to:

$$\lambda \sum (\theta_j - 2\phi\theta_{j-1} + \theta_{j-2})^2,$$

where $\phi = \cos(2\pi d/p)$, where d is the distance between knots, then for high values of

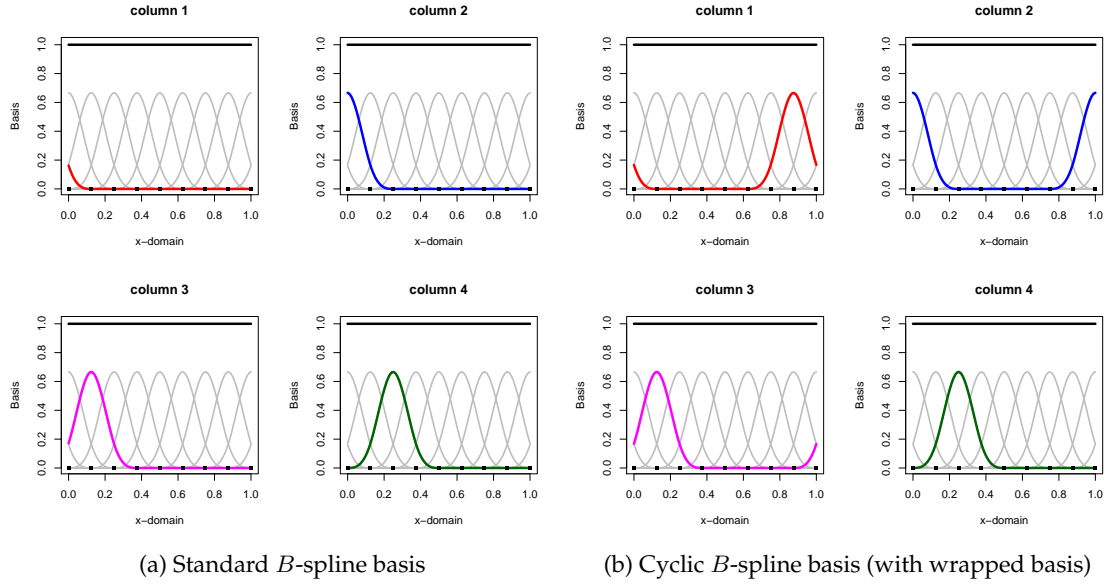


Figure 4.15: Comparison of standard and cyclic cubic B -spline basis. Both figures represent the first four columns of the cyclic B -spline basis.

the smoothing parameter λ , the coefficient vector θ , will tend to a sine function with period p , and gives $\theta_j = a \cos(2\phi j d/p + b)$, where a and b are determined by the data. This forces the fitted periodic data towards a sine signal.

In the spatio-temporal context, an interesting topic for further research is the use of the *harmonic* approach to extend the reduced S-ANOVA model shown in Section 4.4. Considering the *specialized* cyclic basis and penalties for the temporal main effect, $f_t(x_t)$, and combining these bases with the idea of nested B -spline basis. This would lead to consider *nested harmonic B-spline* bases for the spatio-temporal interaction that would also consider a harmonic effect in the interaction term. Following the ideas presented in this Chapter, the construction of an identifiable reduced S-ANOVA model with *harmonic* and *nested harmonic B-spline* bases would require a reparameterization of the model basis into a mixed model, where the fixed unpenalized part is a sine wave function and a non-linear penalized random part.

*"All I want to be is someone that makes
new things and thinks about them".
John Maeda*

Chapter 5

Conclusions and further work

Summary of contributions of the thesis

Statistical methods for the analysis of spatial data have been studied along the years in many different fields of research. Spatial data arises from many disciplines as economics, environmental sciences, climatology, ecology, epidemiology or demography. The different typology and classification of spatial data has also contributed to the development of a wide variety of modelling approaches. The classic classification of spatial data was proposed by [Cressie \(1993\)](#), as: (i) geostatistical data, (ii) areal or regional data and (iii) spatial point patterns. In [Chapter 1](#), we defined the basic characteristics for each type of spatial data and presented illustrative examples. We also provided a review of the classical spatial models used for them.

The main drawback on the classical methods is the strong assumptions needed for consideration (as stationarity or isotropy), that in many situations result unrealistic. These methods also present limitations in estimation procedures, when the spatial data sets are very large. In recent years, there have been an increasing interest in the analysis of spatial data collected across time. The spatio-temporal data modelling has supposed a challenge from the statistical point of view. The incorporation of the temporal dimension has brought an increasing complexity in terms of methodology and computational efficiency.

In this dissertation, we have considered a non-parametric regression approach, based on smoothing techniques for the analysis of spatial and spatio-temporal data. Smoothing techniques have also a long tradition in the analysis of univariate data and density estimation. The interest of applying these techniques to the spatial context lies in studying the observed spatial process as a smooth trend surface with a random noise. We consider the use of low-rank penalized spline regression models (penalized likelihood with B -splines basis) as an unified framework to smooth spatial and spatio-

temporal data. [Chapter 2](#) is devoted to present the main aspects of the methodology for Gaussian and non-Gaussian responses in the generalized linear model (GLM) framework. We consider the P -spline methodology using a mixed model formulation, by the reparameterization of the model basis into a fixed and random effects matrices and the penalty into a block-diagonal matrix. We based the reparameterization on the singular value decomposition of the penalty matrix. The benefits of the mixed model formulation are basically: (i) the estimation of the optimal amount of smoothing is now a problem of estimation of the variance components in a mixed model, and so, the use of standard mixed model estimation procedures (as for example restricted maximum likelihood) and software is available; (ii) the possibility of incorporating complex structures as random effects (spatial effects, correlation structures or longitudinal data) as part of the modelling and estimate them simultaneously to the smoothing; (iii) extend the mixed model methodology to smooth non-Gaussian data in the generalized linear mixed model (GLMM) framework, and use for example penalized quasi-likelihood for estimation. We called this approach, *smoothing mixed models*.

In the case of data arranged in multidimensional grids, the use of the *generalized linear array models* (or GLAM) algorithms developed by [Currie et al. \(2006\)](#) and [Eilers et al. \(2006\)](#) are a great advantage in computational efficiency and storage of large data matrices. The extension of the GLAM methods to the mixed model formulation is also detailed in [Chapter 2](#).

In [Chapter 3](#), we apply the smoothing mixed model approach for scattered or spatial data. Given the structure of the data the GLAM algorithms are not available in this case, [Eilers et al. \(2006\)](#) proposed the use of the row-wise Kronecker product, or simply “*Box-product*”. For this case, the reparameterization into a mixed model is not straightforward, but we demonstrate that using some matrix algebra results the model can be formulated as a mixed model. We illustrated this approach to the different types of spatial data classification.

As part of [Chapter 3](#), in [Lee and Durbán \(2009\)](#), we studied an application to disease mapping. We analyzed the well-known Scottish Lip Cancer and compared several alternatives to deal with overdispersion in spatial count data, as for example a smooth Negative Binomial model. We also presented a new model that combines a smooth model (to account for the large-scale trend variability) and random effects with a *conditional autoregressive* structure (to account for the small-scale local variability), we called this model as *Smooth-CAR*. The simulation study performed also shows us that the Smooth-CAR model performs better in most of the situations where large and/or small scale variability is present. This is a general hybrid-model, so the other proposed alternatives (smooth Poisson, PRIDE a CAR models) are particular cases of this one.

In [Chapter 4](#), we have extended the smoothing mixed models methodology developed in [Chapter 2](#) to the case of including additive models with interactions as in a classical ANOVA model. We called this models *Smooth-ANOVA* models. We have demonstrated how to reparameterize the model bases and penalties to obtain an identifiable model, and also that this procedure is equivalent to impose linear constraints to the original non-transformed coefficients. A small simulation study was carried out to evaluate the performance of the S-ANOVA model in comparison to additive and interaction models. Again, the new model proved to better fit of the simulated data. However, as we have addressed in [Section 4.3](#), the development of inferential aspects an testing for interaction terms is a current topic of special interest for multidimensional S-ANOVA models.

For situations in which we are not interested in modelling a full ANOVA model (a model that includes all main effects and interactions), but we want to consider some of the terms and ignore others (mostly some interactions), we proposed the construction of *reduced S-ANOVA models*. These new models are of special interest in the spatio-temporal case. In [Lee and Durbán \(2010\)](#), we used the reparameterization into a mixed model, and showed how to construct an identifiable model by the correct specification of the model matrices and block-diagonal penalty matrix. We demonstrated that for the reduced S-ANOVA model, the linear constraints can also be obtained in terms of the original non-transformed P -splines coefficients. These constraints are a subset of the linear constraints of the full S-ANOVA model. In the spatio-temporal case, the model matrices can be used in the GLAM framework, thus, array formulation of multidimensional P -spline models yields a unified framework for d -dimensional smoothing. It is possible to represent a d -dimensional $c_1 \times c_2 \times \dots \times c_d$ array of coefficients by Θ , and apply the corresponding constraints. The interpretation of the constraints is also easier using the array form, since they are applied over each of the dimensions of the coefficients array. The array Θ is flattened onto the dimension in which the constraints are applied, and reinstated in vector form.

In practice, it is also easy to extend the model by the incorporation of other relevant covariates as smooth additive terms or as interactions. One of the main benefits of the spatio-temporal reduced S-ANOVA model proposed is the interpretation of the smooth functions and the ability of visualize each of the terms of the decomposition in descriptive plots. The reduced S-ANOVA model also gives a direct interpretation in terms of their smoothing parameters and regression coefficients, since we set independent and separate penalties and coefficients for each smooth term.

With large datasets, the computational implementation of the analyses of spatio-temporal data are very intensive and requires efficient computational methods. In the

P -spline approach, the dimension of the bases involved in the smoothing depends basically on the number of knots, and therefore, the dimensionality of the problem is reduced by setting a moderate number for each covariate dimension. However, when data often present a strong seasonal trend (which is very common in environmental problems), the size of the temporal B -spline basis has to be large (between 20 and 40 equidistant knots) in order to have enough degrees of freedom to capture the temporal structure. In the application of the air pollution ozone levels, we found adequate a number of 4 knots for each of the seven years considered. If a larger sample of monitoring stations would have been considered in the study during a larger time period, the number of parameters in the interaction $B_s \otimes B_t$ could easily be of the order of thousands, and the computational burden prohibitive. Nevertheless, the GLAM methods also have an important role in the algorithms implementation, since they allow us to store the data and model matrices more efficiently and speed up the calculations. This computational aspect is a topic of current research.

As a result of the reduced S-ANOVA model, we considered the use of lower rank B -spline bases for interaction terms in order to avoid the estimation of a large number of parameters. We called this new bases as *nested B -spline bases*. The use of these bases has two important features: (i) the linear constraints necessary to maintain the model identifiability remains the same as using the same marginal bases for the additive terms than for the interactions. This is important in order to keep the hierarchical nature of the S-ANOVA models; (ii) interaction terms constructed with nested B -spline bases achieves a great reduction of the total number of parameters to be estimated, this leads to a more efficient implementation of the estimation algorithms. This approach can also be extended to consider nested B -spline bases for the spatial component, and define the space-time interaction basis as $\tilde{B}_s \otimes \tilde{B}_t$, where \tilde{B}_s is constructed from a reduced set of knots of the marginal basis of x_1 and x_2 . Depending on the spatial data structure, it may be necessary to increase the number of knots in the spatial smooth term, and use a lower dimension basis for the interaction and avoid computational complexity.

Further research

This thesis has shown the usefulness of Penalized splines models in the context of spatial and spatio-temporal data. We have addressed the main aspects related to the construction of identifiable models, based on the reparameterization as a mixed model. Although the connection between smoothing and mixed models is not new, the procedure shown in this thesis using the properties of the singular value decomposition allowed us to extend the P -spline methodology to the context of multidimensional Smooth-

ANOVA models. This S-ANOVA model might be considered as a powerful tool for a wide range of possible problems in many other real applications, for which flexible regression models are required. In general, for those problems where smooth additive and interaction terms are considered as part of the modelling. For instance, as we showed in the analysis of spatial point patterns, we can consider the non-parametric approach as a method for multivariate density estimation. In this context, the S-ANOVA model the idea of consider the multidimensional model in terms of main effects and interactions, can be applied to estimate marginals, joint and conditional density functions using a penalized likelihood Tensor product approach.

In the simulation studies, we have already pointed out the problem of testing several variance components in the context of the smoothing mixed models methodology. The development of computationally efficient tests for several variance components in S-ANOVA models with isotropic interactions, would give a powerful tool to check the adequacy of the models we have proposed in this thesis.

The approach we have developed is based on low-rank models regression, based on the use of GLAM arithmetic. However, in the mixed model formulation, the estimation of the variance components is done by REML and PQL for non-Gaussian responses. As in many optimization problems, there is no global optimization procedure for maximizing the likelihood in high-dimensional data, and thus more research in this direction must be done for fast procedures for estimation of the variance components matrices for the special cases we have considered in this thesis. The study of matrix algebra and computational efficient methods for optimization will be also required to avoid large matrix operations. In the context of GLMM's, [Schall \(1991\)](#) developed an iterative method for the estimation of variance components using REML. The use of this method for the estimation can be easily implemented for the uni-dimensional case or for the estimation of several smoothing parameters in the additive model context. The estimation procedure is done with few iterations of the algorithm. The extension of this type of iterative solutions for models with isotropic interactions would be of great interest and would avoid the use of optimization routines.

From a Bayesian perspective, the use of the P -spline methodology is a topic of current research from the Bayesian community. The Smooth-CAR model can also be considered in this setting, from a hierarchical Bayesian approach. The extension of the S-ANOVA model would be also of interest, since we have shown how to construct identifiable model basis and penalties, the Bayesian approach can be also applied with the proper specification of the stages and the study of the priors on the hyperparameters. The implementation of these models in the Bayesian framework would require the development of fast methods for Gibbs sampling and MCMC methods.

As we have addressed in [Chapter 4](#), the development of the nested B -spline basis have very attractive features for future research, not only in the spatio-temporal context but also in other smoothing problems. In the spatio-temporal context, it can be considered the inclusion of harmonics in the modelling of data with seasonality. If we consider a harmonic B -spline basis and penalty for the temporal main effect, $f_t(\cdot)$, and interesting extension would be to consider nested harmonic B -splines basis and perform the mixed model reparameterization with such harmonic basis and penalties. This would give us identifiable reduced S-ANOVA models with harmonics terms incorporated in the space-time interaction, taking also advantages on the lower rank nested B -spline basis for computational efficiency. We are interested in studying if the reparameterization, or similar ideas can be applied for obtaining the fixed effects with a limiting case of a harmonic term and a random penalized part that also includes a possible harmonic space-time interaction.

In the case of longitudinal studies, using P -splines as mixed models (see [Durbán et al., 2005](#)), the use of nested B -spline bases would be of interest if for example we consider a longitudinal model of form:

$$\mathbf{y}_{ij} = f(t_{ij}) + g_i(t_{ij}) + \epsilon_{ij}, \quad \epsilon \sim \mathcal{N}(0, \sigma^2),$$

for $i = 1, \dots, n$ subjects, observed in $j = 1, \dots, m_i$ times for each subject, f represents the smooth group mean curve and g_i represents the smooth deviation curve for subject i . This model is a generalization of a one-way ANOVA model for sampled curves. An interesting approach is to use a S-ANOVA formulation in a such way that the linear constraints on the coefficients are obtained from the reparameterization of the B -spline basis. The use of nested B -spline bases for the specific curves would also reduced the computational burden.

References

- Aerts, M., Claeskens, G., and Wand, M. P. (2002). Some theory for penalized spline generalized linear models. *J. Stat. Plan. Infer.*, 103:455–470.
- Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60:255–265.
- Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns. *Australian and New Zealand Journal of Statistics*, 42:283–322.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Monographs on Statistics and Applied Probability 101. Chapman & Hall/CRC.
- Bell, M. and Grunwald, G. K. (2004). Mixed models for the analysis of replicated spatial point pattern. *Biostatistics*, 5:633–648.
- Bellman, R. (1961). *Adaptive Control Processes: A guided tour*. Princeton University Press.
- Berke, O. (2004). Exploratory disease mapping: Kriging the spatial risk function from regional count data. *International Journal of Health Geographics*, 3(18):–.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc. B*, 36:192–236.
- Besag, J. (1984). A method for the analysis of field experiments based on first differences. In *Spatial Methods in Field Experiments. Proceedings of the Biometric Society Workshop*, pages 9–11, University of Durham.
- Besag, J. and Green, P. J. (1993). Spatial statistics and bayesian computation. *J. R. Statist. Soc. B*, 55(1):25–37.
- Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82:733–746.

- Besag, J., York, J. C., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43:1–59.
- Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford Statistical Science Series. Oxford University Press, Oxford.
- Bowman, A. W., Giannitrapani, M., and Scott, E. M. (2009). Spatiotemporal smoothing and sulphur dioxide trends over europe. *Applied Statistics*, 58(5):737–752.
- Breslow, N. E. (1984). Extra-poisson variation in log-linear models. *Applied Statistics*, 33(1):38–44.
- Breslow, N. E. and Clayton, D. G. (1993). Aproximated inference in generalised linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.
- Brezger, A., Kneib, T., and Lang, S. (2005). Bayesx: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software*, 14(11):1–22.
- Brezger, A. and Lang, S. (2003). Generalized structured additive regression based on bayesian p-splines. Technical report, Department of Statistics, University of Munich.
- Brezger, A. and Lang, S. (2008). Simultaneous probability statements for bayesian p -splines. *Statistical Modelling*, 8(2):141–168.
- Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Am. Statist. Assoc.*, 93(443):961–994.
- Brumback, B. A., Ruppert, D., and Wand, M. P. (1999). Variable selection and function estimation in additive nonparametric regression using a data-set prior: Comment. *Journal of the American Statistical Association*, 94(447):794–797.
- Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Stat.*, 17:453–555.
- Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Number 30 in Econometric Society Monograph. Cambridge University Press.
- Cantoni, E. and Hastie, T. (2000). Degrees of freedom tests for smoothing splines. Technical report, Dep. Statistics, Stanford University.
- Chen, Z. (1993). Fitting multivariate regression functions by interaction spline models. *J. R. Statist. Soc. B*, 55:473–491.

- Claeskens, G. (2004). Restricted likelihood ratio lack-of-fit tests using mixed spline models. *J. R. Statist. Soc. B*, 66(4):909–926.
- Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009). Asymptotic properties of penalized splines estimators. *Biometrika*, 3:529–544.
- Clayton, D. G. and Kaldor, J. (1987). Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–682.
- Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J. Am. Statist. Assoc.*, 83:596–610.
- Congdon, P. (2006). A model for non-parametric spatially varying regression effects. *Computational Statistics and Data Analysis*, 50:422–445.
- Congdon, P. (2007). Mixtures of spatial and unstructured effects for spatially discontinuous health outcomes. *Computational Statistics and Data Analysis*, 51:3197–3212.
- Coull, B., Schwartz, J., and Wand, M. (2001a). Respiratory health and air pollution: Additive mixed model analyses. *Biostatistics*, 2:337–349.
- Coull, B. A., Ruppert, D., and Wand, M. P. (2001b). Simple incorporation of interactions into additive models. *Biometrics*, 57:539–545.
- Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika*, 92(1):91–103.
- Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. R. Statist. Soc. B*, 66:165–185.
- Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of cross-validation. *Numer. Math.*, 31:377–403.
- Cressie, N. (1993). *Statistics for Spatial Data (Revised Edition)*. John Wiley and Sons, Inc., New York.
- Cressie, N. and Chan, N. H. (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association*, 84(406):393–401.
- Cressie, N. and Huang, H. C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1340.

- Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large data sets. *J. R. Statist. Soc. B*, 70(1):209–226.
- Cressie, N. and Read, T. R. C. (1985). Do sudden infant deaths come in clusters? *Statistics and Decisions*, 2:333–249. Supplement Issue.
- Currie, I. D. and Durbán, M. (2002). Flexible smoothing with P -splines: A unified approach. *Statistical Modelling*, 2:333–349.
- Currie, I. D., Durbán, M., and Eilers, P. H. C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298.
- Currie, I. D., Durbán, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *J. R. Statist. Soc. B*, 68:1–22.
- de Boor, C. (1978). *A practical Guide to Splines*. Springer, Berlin.
- Dean, C., Ugarte, M. D., and Militino, A. F. (2001). Detecting interaction between random region and fixed age effects in disease mapping. *Biometrics*, 57:197–202.
- Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Clarendon Press, Oxford.
- Diggle, P. J. (1981). *Interpreting Multivariate Data*, chapter Some practical methods in the analysis of spatial point patterns, pages 55–73. John Wiley & Sons, New York.
- Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. Chapman & Hall, New York.
- Diggle, P. J. (1985). A kernel method for smoothing point process data. *Applied Statistics*, 34:138–147.
- Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Edward Arnold, London, 2nd edition edition.
- Duchon, J. (1976). *Splines minimizing rotation-invariant semi-norms in Sobolev spaces, constructive theory of functions of several variables*, volume 1 of *Lecture Notes in Mathematics*. Springer.
- Durbán, M. and Currie, I. D. (2003). A note on P -spline additive models with correlated errors. *Computational Statistics*, 18:251–262.
- Durbán, M., Currie, I. D., and Eilers, P. H. C. (2006). Mixed models, array methods and multidimensional density estimation. In *Proceedings of the 21st International Workshop on Statistical Modelling*, pages 143–150, Galway, Ireland.

- Durbán, M., Harezlak, J., Wand, M. P., and Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statist. Med.*, 24:1153–1167.
- EEA (2009). Air pollution by ozone across Europe during summer 2008. Technical Report 2, European Environmental Agency.
- Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis*, 50(1):61–76.
- Eilers, P. H. C., Gampe, J., Marx, B. D., and Rau, R. (2008). Modulation models for seasonal time series and incidence tables. *Statist. Med.*, 27:3430–3441.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. *Stat. Sci.*, 11:89–121.
- Eilers, P. H. C. and Marx, B. D. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, 11(4):758–783.
- Eilers, P. H. C. and Marx, B. D. (2004). Splines, knots and penalties. Technical report.
- Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*, volume 157. CRC Press, second edition edition.
- Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica*, 14:715–745.
- Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random fields prior. *Applied Statistics*, 50:201–220.
- Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modelling (with discussion). *Technometrics*, 1:3–39.
- Fuentes, M., Chen, L., and Davis, J. M. (2008). A class of non-separable and non-stationary spatial temporal covariance functions. *Environmetrics*, 19:487–507.
- Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458):590–600.
- Goodall, C. and Mardia, K. (1994). Challenges in multivariate spatio-temporal modelling. In *In Proceedings of the Seventeenth International Biometric Society*, Hamilton, Ontario, Canada.
- Gössl, C., Auer, D. P., and Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics*, 57(2):554–562.

- Graham, A. (1986). *Kronecker products and matrix calculus: with applications*. Ed. Elis Horwood.
- Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.
- Green, P. J. (1999). Discussion on "The analysis of designed experiments and longitudinal data by smoothing splines" (by Verbyla, A. P., Cullis, B. R., Kenward, M. G. and Welham, S. J.). *J. R. Statist. Soc. C*, 48:304–305.
- Greven, S., Crainiceanu, C., Küchenhoff, H., and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17(4):870–891.
- Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer Series in Statistics. Springer.
- Gu, C. and Wahba, G. (1993). Semiparametric analysis of variance with tensor product thin plate splines. *J. R. Statist. Soc. B*, 55(2):353–368.
- Haggett (1977). *Locational methods in human geography*. Edward Arnold, London.
- Hall, P., Lahiri, S. N., and Polzehl, J. (1995). On bandwidth choice in nonparametric regression with both short- and long-range dependent errors. *Ann. Stat.*, -:1921–1936.
- Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, 92(1):105–118.
- Harville, D. (2000). *Matrix Algebra from a Statistician's Perspective*. Springer.
- Harville, D. A. (1974). Bayesian inference for variance components using error contrasts. *Biometrika*, 61:383–385.
- Hastie, T. (1996). Pseudosplines. *J. R. Statist. Soc. B*, 58:379–396.
- Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1:297–318.
- Hastie, T. and Tibshirani, R. (1987). Generalized additive models: Some applications. *J. Am. Stat. Assoc.*, 82:371–386.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman and Hall, London.

- Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *J. R. Statist. Soc. B*, 55(4):757–796.
- Hinde, J. and Demetrio, C. (1998). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis*, 27:151–170.
- Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. Cambridge: Cambridge University Press.
- Huang, H. C. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis*, 22:159–175.
- Hurvich, C. and Simonoff, J. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *J. R. Statist. Soc. B*, 60:271–293.
- Kammann, E. E. and Wand, M. P. (2003). Geoaddivitive models. *Journal of the Royal Statistical Society, C - Applied Statistics*, 52:1–18.
- Kaüermann, G. (2005). A note on smoothing parameter selection penalized spline smoothing. *Journal of Statistical Planning and Inference*, 127:53–69.
- Kaüermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *J. R. Statist. Soc. B*, 71:487–503.
- Kneib, T. and Fahrmeir, L. (2006). Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, 62:109–118.
- Kostaki, A. and Panousis, V. (2001). Expanding on abridged life table. *Demographic Research*, 5:1–22.
- Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52:119–139.
- Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.
- Laslett, G. M. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications. *J. Am. Stat. Assoc.*, 89:391–409.
- Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics*, 15:209–225.

- Lee, D.-J. and Durbán, M. (2009). Smooth-CAR mixed models for spatial count data. *Computational Statistics and Data Analysis*, 53(8):2958–2979.
- Lee, D.-J. and Durbán, M. (2010). *P*-spline ANOVA-type interaction models for spatio-temporal smoothing. *to appear in Statistical Modelling*.
- Leroux, B. G., Lei, X., and Breslow, N. (1999). *Estimation of disease rates in small areas: a new mixed model for spatial dependence*. Statistical models in Epidemiology, The Environment, and Clinical Trials. Springer, New York.
- Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326.
- Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *J. Roy. Stat. Soc., B*, 61:381–400.
- Liu, S. (1999). Matrix results on the khatri-rao and tracy-singh products. *Linear Algebra and its Applications*, 289:267–277.
- Liu, S. (2002). Several inequalities involving khatri-rao products of positive semidefinite matrices. *Linear Algebra and its Applications*, 354:175–186.
- MacNab, Y. and Dean, C. B. (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*, 57:949–956.
- MacNab, Y. C. and Dean, C. B. (2000). Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models. *Statistics in Medicine*, 19:2421–2435.
- Mardia, K. V. and Goodall, C. (1993). *Spatial-temporal analysis of multivariate environmental monitoring data*, pages 347–386. Elsevier/North Holland, New York, Amsterdam.
- Mardia, K. V., Goodall, C., Redfern, E. J., and Alonso, F. J. (1998). The kriged kalman filter. *Test*, 7:217–285.
- Mardia, K. V., Kent, J. T., Goodall, C. R., and Little, J. A. (1996). Kriging and splines with derivative information. *Biometrika*, 83(1):207–221.
- Marx, B. D. and Eilers, P. C. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, 28:193–209.
- Marx, B. D., Eilers, P. H. C., Gampe, J., and Rau, R. (2010). Bilinear modulation models for seasonal tables of counts. *Statistics and Computing*, 20:191–202.

- Matérn, B. (1986). *Spatial Variation*. Springer-Verlag, lectures notes in statistics edition.
- Matheron, G. (1962). *Traite de Geostatistique Appliquee, Tome I*. Memoires du Bureau de Recherches Geologiques et Minieres. Number 14. Editions Technip, Paris.
- Matheron, G. (1963). *Traite de Geostatistique Appliquee, Tome II: Le Krigeage*. Memoires du Bureau de Recherches Geologiques et Minieres. Number 24. Editions Bureau de Recherche Geologiques et Minieres, Paris.
- McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, 2nd edition.
- McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. John Wiley & Sons, New York.
- Militino, A. F., Ugarte, M. D., and Dean, C. B. (2001). The use of mixture models for identifying high risks in disease. *Statistics in Medicine*, 20(13):2035–2049.
- Nelder, J. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Stat. Soc. A*, 135:370–384.
- Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Am. Statist. Assoc.*, 83:1134–43.
- Nychka, D. (2000). *Spatial-process Estimates as Smoothers*, In Schimek (ed.) *Smoothing and Regression: Approaches, Computation and Applications*. John Wiley & Sons, New York. A contributed chapter to Smoothing and Regression.
- O’Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1:505–527.
- Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.
- Pawitan, Y. (2001). *In All Likelihood : Statistical Modelling and Inference Using Likelihood*. Oxford University Press, USA.
- Perperoglou, A. and Eilers, P. H. C. (2009). Penalized regression and individual deviance effects. *Stat Comput.* DOI: 10.1007/s00180-009-0180-x.
- Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer-Verlag.
- Rao, C. R. and Rao, M. B. (1998). *Matrix Algebra and Its Applications to Statistics and Econometrics*. World Scientific Publishing Company, 1st edition.

- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Stat.*
- Ruppert, D. and Carroll, R. J. (2000). Spatially-adaptive penalties for spline smoothing. *Aust. NZ. J. Stat.*, 42(2):205–223.
- Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Am. Statist. Assoc.*, 90:1257–1270.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, UK. ISBN: 0521785162.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, 3:1193–1256.
- Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–721.
- Scheipl, F., Greven, S., and Küchendorff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics and Data Analysis*, 52(7):3283–3299.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.
- Searle, S., Casella, G., and McCulloch, C. (1992). *Variance components*. Wiley Series in Probability and Mathematical Statistics.
- Self, S. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, 82:605–610.
- Self, S. and Liang, K. Y. (1995). On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *J. R. Statist. Soc. B*, 58:785–796.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall.
- Simonoff, J. (1996). *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer-Verlag, New York.
- Speed, T. (1991). Comment on “BLUP is a good thing: The estimation of random effects”, by Robinson, G.K. *Stat. Sci.*, 6:15–51.

- Stein, M. L. (1999). *Interpolating Spatial Data: Some Theory of Kriging*. Springer-Verlag, New York.
- Stein, M. L. (2005). Space-time covariance functions. *Journal of the American Statistical Association*, 100:310–321.
- Stiratelli, R., Laird, N. M., and Ware, J. H. (1984). Random effects models with serial observations with binary responses. *Biometrics*, 40:719–727.
- Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13(2):689–705.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Stat. Soc. B*, 36:111–147.
- Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects models. *Biometrics*, 50:1171–1177.
- Symons, M. J., Grimson, R. C., and Yuan, Y. C. (1983). Clustering of rare events. *Biometrics*, 39(1):193–205.
- Thurston, S., Wand, M., and Wiencke, J. (2000). Negative binomial additive models. *Biometrics*, 56:139–144.
- Ugarte, M. D., Ibañez, B., and Militino, A. F. (2005). Detection of spatial variation in risk when using car models for smoothing relative risks. *Stochastic Environmental Research and Risk Assessment*, 19:33–40.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for longitudinal data*. Series in statistics. Springer.
- Verbyla, A., Cullis, B., Kenward, M., and Welham, S. (1999). The analysis of designed experiments and longitudinal data using smoothing splines. *J. Roy. Stat. Soc. C*, 48:269–312.
- Wahba, G. (1983). Bayesian “confidence intervals” for the cross-validated smoothing spline. *J. Roy. Stat. Soc. B*, 45:133–150.
- Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia.
- Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995). Smoothing spline anova for exponential families, with application to the winconsin epidemiological study of

- diabetic retinopathy. *The Annals of Statistics*, 23(6):1865–1895. THE 1994 NEYMAN MEMORIAL LECTURE.
- Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.
- Wall, M. M. (2004). A close look at the spatial structure implied by the car and sar models. *Journal of Statistical Planning and Inference*, 121:311–324.
- Waller, L., Carlin, B., Xia, H., and Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92(438):607–617.
- Wand, M. P. (1999). On the optimal amount of smoothing in penalised spline regression. *Biometrika*, 86(4):936–940.
- Wand, M. P. (2002). Vector differential calculus in statistics. *The American Statistician*, 56(1):55–62.
- Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.
- Wang, Y. (1998a). Mixed effects smoothing spline analysis of variance. *J. R. Statist. Soc. B*, 60(1):159–174.
- Wang, Y. (1998b). Smoothing spline models with correlated random errors. *J. Am. Stat. Assoc.*, 93(441):341–348.
- Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5:117–154.
- Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.
- Wood, S. N. (2003). Thin plate regression splines. *J. R. Statist. Soc. B*, 65(1):95–114.
- Wood, S. N. (2006a). *Generalized Additive Models - An introduction with R*. Texts in Statistical Science. Chapman & Hall.
- Wood, S. N. (2006b). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036.
- Yasui, Y. and Lele, S. (1997). A regression method for spatial disease rates: An estimating function approach. *Journal of the American Statistical Association*, 92(437):21–32.

- Zhang, S., Lin, X., Raz, J., and Sowers, M. (1998). Semi-parametric stochastic mixed models for longitudinal data. *J. Am. Stat. Assoc.*, 93:710–719.
- Zheng, P., Durr, P. A., and Diggle, P. J. (2004). Edge-correction for spatial kernel smoothing methods - when is it necessary? In *Proceedings of the GisVet Conference 2004*, Ontario, Canada. University of Guelph.

Appendix A

Appendix to Chapter 2

A.1 Some basic matrix algebra on Kronecker products

See [Harville \(2000\)](#) and [Graham \(1986\)](#).

Let us define the matrices A , B , C and D , such that: $A_{m \times n} = \{a_{ij}\}$, $B_{p \times q} = \{b_{ij}\}$, $C_{n \times u} = \{c_{ij}\}$ and $D_{q \times v} = \{d_{ij}\}$.

Definition A.1 (Kronecker product of two matrices). Kronecker product of two matrices is denoted by $A \otimes B$ and is defined to be the $mp \times nq$ matrix:

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}$$

obtained by replacing each element a_{ij} of A with the $p \times q$ matrix $a_{ij}B$. Thus, the Kronecker product of A and B is a partitioned matrix, comprising m rows and n columns of $p \times q$ dimensional blocks, the ij th of which is $a_{ij}B$.

Lemma A.1. *If k is a scalar, then*

$$(kA) \otimes B = A \otimes (kB) = k(A \otimes B).$$

Lemma A.2 (Distributive property). *The product is distributive with respect to addition, that is*

$$(i) \quad (A + B) \otimes C = A \otimes C + B \otimes C \text{ and}$$

$$(ii) \quad A \otimes (B + C) = A \otimes B + A \otimes C.$$

Lemma A.3 (Associate property). *The product is associative*

$$\mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C}$$

Lemma A.4. *There exists:*

(i) A zero element $\mathbf{0}_{mn} = \mathbf{0}_m \otimes \mathbf{0}_n$

(ii) A unit element $\mathbf{I}_{mn} = \mathbf{I}_m \otimes \mathbf{I}_n$

The unit matrices are all square, for example \mathbf{I}_m in the unit matrix of order $(m \times m)$.

Lemma A.5 (Transpose of the Kronecker product of two matrices). *The transpose of the Kronecker product is the product of the transpose matrices:*

$$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'.$$

Lemma A.6 (Mixed product rule).

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}.$$

One implication of this lemma, is that the product $\mathbf{A} \otimes \mathbf{B}$, can be decomposed as

$$\mathbf{A} \otimes \mathbf{B} = (\mathbf{A} \otimes \mathbf{I}_p)(\mathbf{I}_n \otimes \mathbf{B}) = (\mathbf{I}_m \otimes \mathbf{B})(\mathbf{A} \otimes \mathbf{I}_q).$$

Result in A.6 can be extended (by repeated application) as

$$(\mathbf{A}_1 \otimes \mathbf{B}_1)(\mathbf{A}_2 \otimes \mathbf{B}_2) \cdots (\mathbf{A}_k \otimes \mathbf{B}_k) = (\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_k) \otimes (\mathbf{B}_1 \mathbf{B}_2 \cdots \mathbf{B}_k),$$

where (for $i = 1, 2, \dots, k$) \mathbf{A}_i is a $m_i \times m_{i+1}$ dimensional matrix and \mathbf{B}_i is a $p_i \times p_{i+1}$ dimensional matrix.

Lemma A.7 (Inverse of the Kronecker product of two matrices). *Given $\mathbf{A}_{m \times m}$ and $\mathbf{B}_{n \times n}$ and subject to the existence of the various inverses,*

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}.$$

Using the result in A.6, we have that:

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{A}^{-1} \otimes \mathbf{B}^{-1}) = \mathbf{AA}^{-1} \otimes \mathbf{BB}^{-1} = \mathbf{I}_m \otimes \mathbf{I}_n = \mathbf{I}_{mn}$$

Lemma A.8 (Trace of the Kronecker product of two matrices). *The trace of the Kronecker product of two matrices is*

$$\text{tr}(\mathbf{A} \otimes \mathbf{B}) = \text{tr}(\mathbf{A})\text{tr}(\mathbf{B}).$$

Lemma A.9 (Rank of the Kronecker product of two matrices). *The rank of the Kronecker product of two matrices is*

$$\text{rank}(\mathbf{A} \otimes \mathbf{B}) = \text{rank}(\mathbf{A})\text{rank}(\mathbf{B}).$$

Note that result A.9 implies that the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ has full row/column rank if and only if both \mathbf{A} and \mathbf{B} have full row/column rank. Hence the Kronecker product of \mathbf{A} and \mathbf{B} is non-singular if both \mathbf{A} and \mathbf{B} are non-singular.

Definition A.2 (Kronecker sum). Given matrices $\mathbf{A}_{m \times m}$ and $\mathbf{B}_{n \times n}$, their Kronecker sum denoted by $\mathbf{A} \oplus \mathbf{B}$ is defined as the expression:

$$\mathbf{A} \oplus \mathbf{B} = \mathbf{A} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{B}.$$

Definition A.3 (Matrix direct sum operator). *The matrix direct sum of n matrices, constructs a block-diagonal matrix for a set of square matrices $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_m$, of the form:*

$$\bigoplus_{i=1}^n \mathbf{A}_i = \text{blockdiag}(\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_m) = \begin{pmatrix} \mathbf{A}_1 & & & \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ & & & \mathbf{A}_m \end{pmatrix},$$

Remark A.1. *For any diagonal matrix $\mathbf{A} = \{a_i\}$ of order m , and the identity matrix \mathbf{I} of order n . We define:*

$$\begin{aligned} \mathbf{A}_m \otimes \mathbf{I}_n &= \bigoplus_{i=1}^m a_i \mathbf{I}_n = \text{blockdiag}(a_1 \mathbf{I}_n, a_2 \mathbf{I}_n, \dots, a_m \mathbf{I}_n), \\ \mathbf{I}_n \otimes \mathbf{A}_m &= \bigoplus_{j=1}^n \mathbf{A}_m = \text{blockdiag}(\underbrace{a_1, \dots, a_m}_1, \underbrace{a_1, \dots, a_m}_2, \dots, \underbrace{a_1, \dots, a_m}_n). \end{aligned}$$

A.2 Array methods

A.2.1 Basic array arithmetic

In this section we introduced some notation and definitions of array methods proposed in [Currie et al. \(2006\)](#) and [Eilers et al. \(2006\)](#) and the extensions in d -dimensions. It is also considered the computational details in R.

Definition A.4 (Row tensor). The row tensor of a matrix \mathbf{X} with c columns is defined as

$$\mathcal{G}(\mathbf{X}) = (\mathbf{X} \otimes \mathbf{1}') \odot (\mathbf{1}' \otimes \mathbf{X})$$

where $\mathbf{1}$ is a vector of 1's of length c , and \odot is the element-by-element product or Hadamard product. The `Rten.r` function in R implements this:

```
# Row tensor of a matrix X

Rten <- function(X){
  one <- matrix(1, 1, ncol(X))
  kronecker(X, one) * kronecker(one, X)
}
```

Definition A.5 (Row tensor of two matrices). We can extend the [A.4](#) to the row tensor of matrices $\mathbf{X}_1, n \times c_1$ and $\mathbf{X}_2, n \times c_2$, or row-wise Kronecker of *Box-product* defined as

$$\mathcal{G}(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{X}_1 \otimes \mathbf{1}'_{c_2}) \odot (\mathbf{1}'_{c_1} \otimes \mathbf{X}_2)$$

where $\mathbf{1}_{c_1}$ and $\mathbf{1}_{c_2}$ are vectors of 1's of length c_1 and c_2 respectively. Note that $\mathcal{G}(\mathbf{X}, \mathbf{X}) = \mathcal{G}(\mathbf{X})$.

```
Rten2 <- function(X1,X2){
  one.1 <- t(rep(1,ncol(X1)))
  one.2 <- t(rep(1,ncol(X2)))
  kronecker(X1, one.1) * kronecker(one.2, X2)
}
```

Definition A.6 (\mathcal{H} -transform). The \mathcal{H} -transform of the d -dimensional array \mathbf{A} of size $c_1 \times c_2 \times \dots \times c_d$ by the matrix \mathbf{X} of size $r \times c_1$ is denoted $\mathcal{H}(\mathbf{X}, \mathbf{A})$ and defined as: let \mathbf{A}^* of size $c_1 \times c_2 c_3 \dots c_d$ the matrix obtained by flattening dimensions 2 to d of \mathbf{A} ; form

the matrix product $\mathbf{X}\mathbf{A}^*$ of size $r \times c_2 c_3 \dots c_d$; then $\mathcal{H}(\mathbf{X}, \mathbf{A})$ is the d -dimensional array of size $r \times c_2 \times \dots \times c_d$ obtained from $\mathbf{X}\mathbf{A}^*$ by reinstating dimensions 2 to d of \mathbf{A} .

In one dimension $\mathbf{A} = \mathbf{a}$, so $\mathcal{H}(\mathbf{X}, \mathbf{a})$, while in two dimensions $\mathcal{H}(\mathbf{X}, \mathbf{A}) = \mathbf{X}\mathbf{A}$. Thus the \mathcal{H} -transform generalizes premultiplication of vectors and matrices by a matrix.

```
# H-transform of an array A by a matrix X
# i.e. multiply a matrix onto an array
# and output an array
H <- function(X, A){
  d <- dim(A)
  M <- matrix(A, nrow = d[1])
  XM <- X %*% M
  array(XM, c(nrow(XM), d[-1]))
}
```

Definition A.7 (Array rotation). We need to generalize the transpose of a matrix \mathbf{A} . The rotation of the d -dimensional array \mathbf{A} of size $c_1 \times c_2 \dots c_d$ is the d -dimensional array $R(\mathbf{A})$ of size $c_2 \times c_3 \dots c_d \times c_1$ obtained by permuting the indices of \mathbf{A} .

```
# Rotation of an array A i.e. transpose an array
Rotate = function(A){
  d = 1:length(dim(A))
  d1 = c(d[-1], d[1])
  aperm(A, d1)
}
```

[A.7](#) and [A.6](#) can be conveniently combined in:

Definition A.8 (Rotated \mathcal{H} -transform). The rotated \mathcal{H} -transform of the array \mathbf{A} by the matrix \mathbf{X} is given by

$$\rho(\mathbf{X}, \mathbf{A}) = \mathcal{R}(\mathcal{H}(\mathbf{X}, \mathbf{A}))$$

```
# Rotated H-transform of an array A by a matrix X
# i.e. multiply a matrix onto an array,
# convert to an array and transpose the result
RH <- function(X, A){
  Rotate(H(X, A))
}
```

A.2.2 GLAM algebraic operations

In order to illustrate the GLAM arithmetic effectiveness, let us suppose the d -dimensional case. Let \mathbf{Y} be the $n_1 \times n_2 \times \dots \times n_d$ data array and $\mathbf{y} = \text{vec}(\mathbf{Y})$ be the vector equivalent of length $n_1 n_2 \dots n_d \times 1$. Given the individual B -splines basis $\mathbf{B}_1, \dots, \mathbf{B}_d$ of dimensions $n_1 \times c_1$ to $n_d \times c_d$. The model basis is given by:

$$\mathbf{B} = \mathbf{B}_d \otimes \dots \otimes \mathbf{B}_2 \otimes \mathbf{B}_1, \quad (\text{A.1})$$

of dimensions $n_1 n_2 \dots n_d \times c_1 c_2 \dots c_d$. Given the coefficient array Θ and its vector coefficient θ , of dimensions $c_1 \times c_2 \times \dots \times c_d$ and $c_1 c_2 \dots c_d \times 1$ respectively. The idea is to transform the array Θ successively by the marginal model matrices \mathbf{B}_i , $i = 1, 2, \dots, d$. For this we need to define premultiplication of d -dimensional arrays, such as Θ , by a matrix. The basic operations in a GLAM are: (i) *Linear functions*, (ii) *inner products*, and (iii) *diagonal functions*.

Definition A.9 (Linear functions or \mathcal{K} -form). Linear functions involve the computation of matrix-by-vector products as $\mathbf{B}\theta$ and $\mathbf{B}\mathbf{W}\mathbf{z}$. For a d -dimensional array, this operation can be using GLAM methods as:

$$\rho(\mathbf{B}_d, \dots, \rho(\mathbf{B}_2, \rho(\mathbf{B}_1, \Theta))),$$

with $\rho(\cdot)$ in A.8. In R, suppose the $3d$ case, we have:

```
# Linear function:
GLAM <- RH(B3, RH(B2, RH(B1, Theta)))
# the result is of dim. n1 x n2 x n3
```

Similarly, for non-Gaussian data, with matrix \mathbf{W} of weights and *working vector* \mathbf{z} , we have for $\mathbf{B}\mathbf{W}\mathbf{z}$, i.e. $\rho(\mathbf{B}_d, \dots, \rho(\mathbf{B}_2, \rho(\mathbf{B}_1, \mathbf{W}\mathbf{z})))$.

Definition A.10 (Inner products or \mathcal{A} -forms). The elements of the inner product $\mathbf{B}'\mathbf{W}\mathbf{B}$ are given by the d -dimensional array:

$$\rho(G(\mathbf{B}_d)', \dots, \rho(G(\mathbf{B}_2)', \rho(G(\mathbf{B}_1)', \mathbf{W})))$$

This result is a $c_1^2 \times c_2^2 \times \dots \times c_d^2$ array which must be rearranged into the square matrix $\mathbf{B}'\mathbf{W}\mathbf{B}$ of size $c_1 c_2 \dots c_d \times c_1 c_2 \dots c_d$. In R, suppose the $3d$ case, we have:

```

# Inner Product:
GLAM <- RH(t(Rten(B3)), RH(t(Rten(B2)), RH(t(Rten(B1)), W)))
dim(GLAM) <- c(c1,c1,c2,c2,c3,c3)
GLAMaux <- aperm(GLAM, c(1,3,5,2,4,6))
# Rearranged the GLAM array c1^2 x c2^2 x c3^2 into a
# square matrix c1c2c3 x c1c2c3
GLAM <- matrix(GLAMaux, nrow = c1 * c2 * c3)

```

Definition A.11 (Diagonal functions). Let \mathbf{S}_m , a square matrix of dimensions $c_1 c_2 c_3 \times c_1 c_2 c_3$, and let \mathbf{S} the d -dimensional array, $c_1^2 \times c_2^2 \times c_3^2$ of the reorganized elements of \mathbf{S}_m . The diagonal elements of $\text{Var}(\mathbf{B}\hat{\boldsymbol{\theta}})$ are obtained by setting \mathbf{S}_m equal to $(\mathbf{B}'\hat{\mathbf{W}}_\delta\mathbf{B})^{-1}$. It is easy to show in d -dimensions:

$$\rho(G(\mathbf{B}_d), \dots, \rho(G(\mathbf{B}_2), \mathbf{S})).$$

A.2.3 GLAM as mixed models

In this section, we adapt the GLAM operations in Section A.2.2 to the Kronecker products involved in the mixed model matrices as shown in Section 2.3.2.

Definition A.12 (\mathcal{A}_1 -form). Given the inner product of the form $\mathbf{X}'\mathbf{W}\mathbf{X}$ and the d -dimensional array $\mathbf{X} = \mathbf{X}_d \otimes \dots \otimes \mathbf{X}_1$ of dimensions $n_i \times c_i$, for $i = 1, \dots, d$. In $2d$:

$$(\mathbf{X}_2 \otimes \mathbf{X}_1)'(\mathbf{X}_2 \otimes \mathbf{X}_1) = \mathbf{X}_2' \mathbf{X}_2 \otimes \mathbf{X}_1' \mathbf{X}_1 \quad (\text{A.2})$$

The \mathcal{A}_1 -form is

$$\rho(G(\mathbf{X}_2)', \rho(G(\mathbf{X}_1)', \mathbf{W})) \quad (\text{A.3})$$

of dimensions $c_1^2 \times c_2^2$ and where \mathbf{W} is a $n_1 \times n_2$ matrix of ones, i.e. $\mathbf{W} = \mathbf{1}\mathbf{1}'$.

```

# A1-form as in Eq. 2.6

A1.form<-function(A,B) {
  n2= nrow(A); n1=nrow(B)
  c2= ncol(A); c1=ncol(B)
  M <- matrix(rep(1,n1*n2),nrow=n1)
  Fast<-RH(t(Rten(A)), RH(t(Rten(B)), M))
  # Rearrangement of the array
  dim(Fast)<-c(c1,c1,c2,c2)
  Fast1<-aperm(Fast,c(1,3,2,4))
  Fast <- matrix(Fast1,nrow=c1*c2)
  return(Fast)
}

```

Definition A.13 (A_2 -form). Given the inner product of the form $\mathbf{X}'\mathbf{W}\mathbf{Z}$ and the d -dimensional array $\mathbf{X} = \mathbf{X}_d \otimes \dots \otimes \mathbf{X}_1$ of dimensions $n_i \times c_i$, for $i = 1, \dots, d$ and $\mathbf{Z} = \mathbf{Z}_d \otimes \dots \otimes \mathbf{Z}_1$ of dimensions $n_i \times d_i$. In 2d:

$$(\mathbf{X}_2 \otimes \mathbf{X}_1)'(\mathbf{Z}_2 \otimes \mathbf{Z}_1) = \mathbf{X}_2' \mathbf{Z}_2 \otimes \mathbf{X}_1' \mathbf{Z}_1 \quad (\text{A.4})$$

The A_2 -form is

$$\rho(G(\mathbf{Z}_2, \mathbf{X}_2)', \rho(G(\mathbf{Z}_1, \mathbf{X}_1)', \mathbf{W})) \quad (\text{A.5})$$

of dimensions $c_1^2 \times c_2^2$ and where \mathbf{W} is a $n_1 \times n_2$ matrix of ones, i.e. $\mathbf{W} = \mathbf{1}\mathbf{1}'$.

```

# A2-form

A2.form<-function(A,B,C,D) {
  n2=nrow(A); c2=ncol(C); d1=ncol(B)
  n1=nrow(B); c1=ncol(D); d2=ncol(A)
  M <- matrix(rep(1,n1*n2),nrow=n1)
  Fast<-RH(t(Rten2(C,A)), RH(t(Rten2(D,B)), M))
  # Rearrangement of the array:
  dim(Fast)<-c(d1,c1,d2,c2)
  Fast1<-aperm(Fast,c(1,3,2,4))
  Fast <- matrix(Fast1,nrow=ncol(B)*ncol(A))
  return(Fast)
}

```

A.3 Software considerations

There exists several R packages available for the implementation of the Penalized splines methodology. In this Section, we present the basic R code to construct the smoothing mixed model basis and and it usage in the standard mixed model packages in R.

Definition A.14 (R-function `bspline.r`). This function computes the B -spline regression basis for covariate x . The user have to specify the number of intervals (`ndx`) in the x -domain and the degree of the B -splines (`bdeg`), usually a cubic spline. This function requires to load the `splines` library, with command `library(splines)`.

```
bspline<-function(x,ndx,bdeg){
  xmin<-min(x); xmax<-max(x)
  xmax <- xmax + 0.01*(xmax-xmin); xmin <- xmax - 0.01*(xmax-xmin)
  dx <- (xr - xl)/ndx
  knots <- seq(xl - bdeg*dx, xr + bdeg*dx, by=dx)
  B <- spline.des(knots, x, bdeg+1, 0*x)$design
  B
}
```

Definition A.15 (R-function `MMbasis.r`). This function computes the basic elements of the mixed model reparameterization shown in Section 2.2.1. The `MMbasis.r` function, has includes an additional argument `pord`, with respect to `bspline.r`, for the penalty order (q). This functions has as outputs: (i) the mixed model matrices: X and Z ; (ii) the vector d are the non-zero eigenvalues of the SVD of the penalty matrix P , i.e. $\tilde{\Sigma}$. Additional elements as the B -spline basis (Z), and the matrix of q differences (D) can also be obtained as outputs.

```
MM.basis<-function(x,ndx,bdeg,pord){
  B<-bspline(x,ndx,bdeg); c<-ncol(B)
  D<-diff(diag(c),differences=pord)
  P<-t(D)%*%D
  P.svd=svd(P)
  Us<-(P.svd$u)[,1:(c-pord)]
  d<-(P.svd$d)[1:(c-pord)]
  Z<-B%*%Us
  X<-NULL
  for(i in 0:(pord-1)){X=cbind(X,x^i)}
  output<-list(X=X,Z=Z,d=d,B=B,c=c,D=D,P=P)
  return(output)
}
```

A.3.1 Function `lme()` in `nlme` R package

For constructed model matrices X and Z , it is possible to use the capabilities of the `lme` function in `nlme` package as follows. Note that, function `lme()`, requires a grouped data structure in order to fit the model and use a random effects matrix Z . Now, we illustrate an example in one dimension, of how to fit a P -spline fit using the mixed model reparameterization shown through Section 2.2.1.

```
library(nlme)
n <- length(y)
# create a dummy grouping variable
# equal to 1 indicates, no nested data
Id<- factor(rep(1,n))
```

The `pdIdent()` function is used to construct the random effects matrix Z , however we need to specify that the covariance matrix G should be a multiple of the identity matrix. Therefore, a simple solution is to multiply the random effect matrix Z by $\tilde{\Sigma}^{-0.5}$, i.e.

$$Z^* = Z\tilde{\Sigma}^{-0.5},$$

where $Z = BU_s$ as defined in (2.38). Then, including Z^* as the random effects matrix in `lme()`, we have that a covariance matrix equal to $G^* = \sigma_\alpha^2 I$.

```
Z.star <- Z%*%diag(1/sqrt(d))

Model.Mat<-data.frame(y,X,Z.star)

# specify a pdIdent class for random
# effects
Z.block<-list(list(Id=pdIdent(~Z.star-1)))
Z.block<-unlist(Z.block,recursive=FALSE)
```

```
# create a grouped data object for
# fixed effects
DataFrame<-groupedData(y~X[, -1]|Id, data=Model.Mat)

# run lme
fit<-lme(y~X[, -1], data=DataFrame, random=Z.block)
```

The estimates of the coefficients β and α and the variance components σ^2 and σ_α^2

can be obtained by:

```
# Coefficients
beta.hat<-fit$coef$fixed
alpha.hat<-as.vector(unlist(fit$coef$random))

# Variance components
s2.<- fit$sigma^2
s2.alpha<- s2.*exp(2*unlist(fit$modelStruct))
```

The fitted curve, \hat{f} , can be obtained as

```
f.hat <- fit$fitted[,2] # the 2nd column
```

Definition A.16 (R-function `Conf.Bands.r`). It is also possible to obtain the confidence bands, as in [Ruppert et al. \(2003\)](#), with the function `Conf.Bands.r`. For given values of the variance components `s2.alpha` and `s2.alpha`, and fitted curve `f.hat`. We obtain as output the lower and upper confidence bands of the fitted curve.

```
Conf.Bands<-function (X,Z,f.hat,s2.,s2.alpha){
  C<-cbind(X,Z)
  lambda<-(sigma2/sigma2.alpha)
  D=diag(c(rep(0,ncol(X)),rep(lambda,ncol(Z))))
  S=sigma2*rowSums(C%%solve(t(C)%*%C+D)*C)
  CB.lower=f.hat-1.96*sqrt(S)
  CB.upper=f.hat+1.96*sqrt(S)
  CB=cbind(CB.lower,CB.upper)
  CB
}
```

It is also possible to obtain the AIC and BIC values and the effective degrees of freedom, from the fitted model by `lme()`, as follows

```
# Trace of the Hat-matrix, but it avoids the
# direct calculation of H
EDF=sum(hat(fit$qr)[1:n])

# Deviance for Gaussian data
Dev <- sum(fit$residuals[1:n]^2)
AIC <- Dev + 2 * EDF
BIC <- Dev + log(n) * EDF
```


The `lme` function incorporates more useful structures for practical modelling, as for example for longitudinal data or correlated data. The book by [Pinheiro and Bates \(2000\)](#) is a very useful monography for mixed effects models. However, the main limitation of the `lme` function is the extension to multidimensional case when an anisotropic smoothing is considered. The specification of the covariance structure of the random effects, must be specified as a multiple of a identity matrix. As we showed in Section 2.3.2, the penalty matrix F in (2.70) in our approach, includes de anisotropic case with different smoothing parameters for each dimension. Therefore, the `pdIdent` class is not possible to use unless we consider an isotropic model, with an unique smoothing parameter λ . We will discuss in subsection 4.2.1 the use of `lme` in the additive models case ([Hastie and Tibshirani, 1990](#)), of P -splines models of the form

$$f(x_1) + f(x_2) + \dots + f(x_d),$$

where there is no interaction among covariates, and therefore each smooth function has its own smoothing parameter.

A.3.2 Function `glmmPQL()` in MASS R package

For fitting generalized linear mixed models (GLMM's), package `MASS` includes the function `glmmPQL`, that uses PQL for estimation by successive calls to the `lme` function. The syntax is therefore similar to `lme`. For example, for Poisson data and model matrices defined as above, we use:

```
library(MASS)
fit.Pglmm<-glmmPQL(y~X[, -1], data=DataFrame,
                    random=Z.block, family=poisson)
```

Appendix B

Appendix to Chapter 3

B.1 On matrix algebra of Khatri-Rao and Tracy-Singh products

Consider matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{C} = (c_{ij})$ of order $m \times n$ and $\mathbf{B} = (b_{kl})$ of order $p \times q$. The Kronecker product of two matrices was defined in [A.1](#). In this appendix, we define some useful definitions on matrix products to demonstrate the mixed model reparameterization in spatial data.

Definition B.1 (Hadamard Product). The Hadamard product or *element-wise matrix product*, denoted by symbol \odot is defined as:

$$\mathbf{A} \odot \mathbf{C} = (a_{ij}c_{ij})_{ij} = \mathbf{C} \odot \mathbf{A},$$

where $a_{ij}c_{ij}$ is the ij^{th} scalar element and $\mathbf{A} \odot \mathbf{B}$ is of order $m \times n$.

Lemma B.1 (Properties of the Hadamard product). *Let \mathbf{A} , \mathbf{B} , \mathbf{C} and \mathbf{D} be matrices of the same order, then*

(i) *Commutative property:* $\mathbf{A} \odot \mathbf{B} = \mathbf{B} \odot \mathbf{A}$.

(ii) *Distributive property:* $\mathbf{A} \odot (\mathbf{B} \odot \mathbf{C}) = (\mathbf{A} \odot \mathbf{B}) \odot \mathbf{C}$

(iii) *Associative property:* $(\mathbf{A} + \mathbf{B}) \odot (\mathbf{C} + \mathbf{D}) = \mathbf{A} \odot \mathbf{C} + \mathbf{A} \odot \mathbf{D} + \mathbf{B} \odot \mathbf{C} + \mathbf{B} \odot \mathbf{D}$

(iv) *If $m = n$, $\mathbf{A} = (a_{ij})$ is any matrix, and \mathbf{I}_m is the identity matrix, then*

$$\mathbf{A} \odot \mathbf{I}_m = \text{diag}(a_{11}, a_{22}, \dots, a_{mm}).$$

(v) *Transpose:* $(\mathbf{A} \odot \mathbf{B})' = \mathbf{A}' \odot \mathbf{B}'$

Consider matrices \mathbf{A} and \mathbf{C} of order $m \times n$ and \mathbf{B} of order $p \times q$. Let $\mathbf{A} = (A_{ij})$ be partitioned with A_{ij} of order $m_i \times n_j$ as the ij^{th} block submatrix, $\mathbf{C} = (C_{ij})$ be partitioned with C_{ij} of order $m_i \times n_j$ as the ij^{th} submatrix and $\mathbf{B} = (B_{kl})$ be partitioned with B_{kl} of order $p_k \times q_l$ as the $(k, l)^{th}$ block submatrix. Then

$$\sum_{i=1}^r m_i = m ; \quad \sum_{j=1}^s n_j = n ; \quad \sum_{k=1}^t p_k = p ; \quad \sum_{l=1}^h q_l = q$$

We have the next definitions:

Definition B.2 (Tracy-Singh Product). The Khatri-Rao product of two matrices (denoted by symbol \circ) is:

$$\mathbf{A} \circ \mathbf{B} = (A_{ij} \circ B)_{ij} = ((A_{ij} \otimes B_{kl})_{kl})_{ij}$$

where A_{ij} is the ij^{th} submatrix of order $m_i \times n_j$ and B_{kl} is the kl^{th} submatrix of order $p_k \times q_l$, $A_{ij} \odot \mathbf{B}$ is the ij^{th} submatrix of order $m_i p \times n_j q$.

Then $A_{ij} \otimes B_{kl}$ is the kl^{th} submatrix of order $m_i p_k \times n_j q_l$ and $(A_{ij} \circ B)$ of order $m_i p \times n_j q$ and $\mathbf{A} \circ \mathbf{B}$ of order $mp \times nq$.

Lemma B.2. For a non-partitioned matrix \mathbf{A} , their $\mathbf{A} \circ \mathbf{B}$ is $\mathbf{A} \otimes \mathbf{B}$, i.e. for $\mathbf{A} = (a_{ij})$, where a_{ij} is scalar, we have

$$\begin{aligned} \mathbf{A} \circ \mathbf{B} &= (a_{ij} \circ \mathbf{B})_{ij} \\ &= ((a_{ij} \otimes B_{kl})_{kl})_{ij} \\ &= ((a_{ij} B_{kl})_{kl})_{ij} \\ &= (a_{ij} \mathbf{B})_{ij} = \mathbf{A} \otimes \mathbf{B} \end{aligned}$$

Lemma B.3. For column-wise partitioned \mathbf{A} and \mathbf{B} , their $\mathbf{A} \circ \mathbf{B}$ is $\mathbf{A} \otimes \mathbf{B}$.

Definition B.3 (Khatri-Rao Product). The Khatri-Rao Product, denoted by symbol $*$, is also called “column-wise” Kronecker product, and defined as:

$$\mathbf{A} * \mathbf{B} = (A_{ij} \otimes B_{ij})_{ij},$$

where $\mathbf{A} = [A_{ij}]$ and $\mathbf{B} = [B_{kl}]$ are partitioned matrices of order $m \times n$ and $p \times q$, respectively, A_{ij} is of order $m_i \times n_j$, B_{kl} of order $p_k \times q_l$, $A_{ij} \otimes B_{ij}$ of order $m_i p_k \times n_j q_l$. Given $\mathbf{A} \in \mathbb{R}^{I \times K}$ and $\mathbf{B} \in \mathbb{R}^{J \times K}$, $\mathbf{A} * \mathbf{B}$ is $(IJ) \times K$.

$$\mathbf{A} * \mathbf{B} = [a_{:1} \otimes b_{:1} \quad : \quad a_{:2} \otimes b_{:2} \quad : \quad \dots \quad : \quad a_{:k} \otimes b_{:k}]$$

Observe that the matrices in a Khatri-Rao product all have the same number of columns. Furthermore, if \mathbf{a} and \mathbf{b} are vectors, then the Khatri-Rao and Kronecker products are identical, i.e., $\mathbf{a} \otimes \mathbf{b} = \mathbf{a} * \mathbf{b}$. The Khatri-Rao product has properties that involve the Hadamard product,

Proposition B.1 (Properties of Khatri-Rao product). *Let $\mathbf{A} \in \mathbb{R}^{I \times L}$, $\mathbf{B} \in \mathbb{R}^{J \times L}$, $\mathbf{C} \in \mathbb{R}^{K \times L}$. Then*

- $\mathbf{A} * \mathbf{B} * \mathbf{C} = (\mathbf{A} * \mathbf{B}) * \mathbf{C} = \mathbf{A} * (\mathbf{B} * \mathbf{C})$
- $(\mathbf{A} * \mathbf{B})'(\mathbf{A} * \mathbf{B}) = \mathbf{A}'\mathbf{A} \odot \mathbf{B}'\mathbf{B}$ and
- $(\mathbf{A} * \mathbf{B})^\dagger = ((\mathbf{A}'\mathbf{A}) \odot (\mathbf{B}'\mathbf{B}))^\dagger (\mathbf{A} * \mathbf{B})'$

Lemma B.4. *For a non-partitioned matrix \mathbf{A} , their $\mathbf{A} * \mathbf{B}$ is $\mathbf{A} \otimes \mathbf{B}$, i.e. for $\mathbf{A} = (a_{ij})$, where a_{ij} is a scalar, we have*

$$\mathbf{A} * \mathbf{B} = (a_{ij} \otimes \mathbf{B}_{ij})_{ij} = (a_{ij} \mathbf{B})_{ij} = \mathbf{A} \otimes \mathbf{B}.$$

Lemma B.5. *For non-partitioned matrices \mathbf{A} and \mathbf{B} , their $\mathbf{A} * \mathbf{B}$ is $\mathbf{A} \circ \mathbf{B}$, i.e. for $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$, where a_{ij} and b_{ij} are scalar, we have,*

$$\mathbf{A} * \mathbf{B} = (a_{ij} \otimes b_{ij})_{ij} = (a_{ij} b_{ij})_{ij} = \mathbf{A} \circ \mathbf{B},$$

where $\mathbf{A}_{ij} \otimes \mathbf{B}_{ij}$ is of order $m_i p_i \times n_j q_j$ and $\mathbf{A} * \mathbf{B}$ of order $(\sum m_i p_i) \times (\sum n_j q_j)$.

Theorem B.1. *Let $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}$ and \mathbf{F} be compatibility partitioned matrices, then*

- (i) $(\mathbf{A} \circ \mathbf{B})(\mathbf{D} \circ \mathbf{E}) = (\mathbf{AD}) \circ (\mathbf{BE})$ if \mathbf{AD} and \mathbf{BE} are well defined.
- (ii) $(\mathbf{A} \circ \mathbf{B})^+ = \mathbf{A}^+ \circ \mathbf{B}^+$ for the Moore-Penrose Inverse
- (iii) $\mathbf{A} * \mathbf{B} \neq \mathbf{B} * \mathbf{A}$ in general
- (iv) $\mathbf{C} * \mathbf{B} = \mathbf{C} * \mathbf{B}$ where $\mathbf{C} = (c_{ij})$ and c_{ij} is a scalar
- (v) $(\mathbf{A} \circ \mathbf{B})' = \mathbf{A}' \circ \mathbf{B}'$
- (vi) $(\mathbf{A} + \mathbf{D}) * (\mathbf{B} + \mathbf{E}) = \mathbf{A} * \mathbf{B} + \mathbf{A} * \mathbf{E} + \mathbf{D} * \mathbf{B} + \mathbf{D} * \mathbf{E}$
- (vii) $(\mathbf{A} * \mathbf{B}) * \mathbf{F} = \mathbf{A} * (\mathbf{B} * \mathbf{F})$
- (viii) $(\mathbf{A} * \mathbf{B}) \odot (\mathbf{D} * \mathbf{E}) = (\mathbf{A} \odot \mathbf{D}) * (\mathbf{B} \odot \mathbf{E})$

Proposition B.2. Let A, B, C and D be four matrices of orders $p \times n$, $m \times n$, $m \times p$, and $n \times m$, respectively. Then

$$(C \otimes D)(A * B) = (CA) * (DB),$$

Proof. Let $a_{:i}$ be the i^{th} column of A and $b_{:i}$, the i^{th} column of B , $i = 1, 2, \dots, n$. Then the i^{th} column of CA is $Ca_{:i}$ and that of DB is $Db_{:i}$. Consequently, the i^{th} column of $(CA) * (DB)$ is $Ca_{:i} \otimes Db_{:i} = (C \otimes D)(a_{:i} \otimes b_{:i})$ which is precisely the i^{th} column of $(C \otimes D)(A * B)$. ■

Proposition B.3. Let A and B be two matrices of orders $p \times n$, and $m \times n$, respectively. Then

$$(A * B) = (A' \square B')'.$$

Proof. See Theorem 1 in [Liu \(1999\)](#). ■

Definition B.4 (Box Product \square). or “row-wise” Kronecker Product. Let A and B be two matrices of orders $m \times p$, and $m \times q$,

$$A \square B = (A \otimes \mathbf{1}'_B) \odot (\mathbf{1}'_A \otimes B),$$

of dimension $m \times pq$.

Theorem B.2. Let $A \in \mathbb{R}^{I \times L}$, $B \in \mathbb{R}^{J \times L}$ and $C \in \mathbb{R}^{K \times L}$.

$$(i) \quad (A * B)' = A' \square B'$$

$$(ii) \quad (A \square B)' = A' * B'$$

$$(iii) \quad (A * B) = (A' \square B')'$$

$$(iv) \quad (A \square B) = (A' * B')'$$

Proof. Immediate by definition of Khatri-Rao ($*$) and row-wise or Box-product (\square). ■

Proposition B.4. Let A, B, C and D be four matrices of orders $p \times n$, $m \times n$, $m \times p$, and $n \times m$, respectively. Then:

$$\begin{aligned} (C \otimes D)(A' \square B')' &= ((CA)' \square (DB)')' \\ &= (A' C' \square B' D')'. \end{aligned}$$

Taking the transpose in both sides, we have

$$(A' \square B')(C \otimes D)' = (A' C' \square B' D')$$

and finally given [A.5](#)

$$(A' \square B')(C' \otimes D') = (A' C' \square B' D').$$

Proof. Immediate by result (iii) in [Theorem B.2](#) and [B.2](#). ■

Smoothing mixed models for spatial and spatio-temporal data

Autor: Dae-Jin Lee

Directora: María L. Durbán Reguera

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Vocal:

Vocal:

Secretario:

Calificación:

Leganés, de de